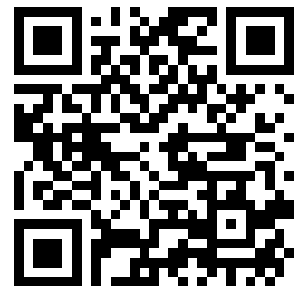

This is a reproduction of a library book that was digitized by Google as part of an ongoing effort to preserve the information in books and make it universally accessible.

GoogleTM books

<https://books.google.com>



WEAPONS SYSTEMS FUNDAMENTALS



**Synthesis
of
Systems**

**PUBLISHED BY DIRECTION OF
THE CHIEF OF THE BUREAU OF NAVAL WEAPONS**

NAVWEPS OP 3000

VOLUME 3

UNIVERSITY OF
ARIZONA LIBRARY
Documents Collection
FEB 3 1964

UNIVERSITY OF MICHIGAN



3 9015 10495 6142

NAVWEPS OP 3000 (VOLUME 3)

**WEAPONS SYSTEMS
FUNDAMENTALS**

**SYNTHESIS
OF
SYSTEMS**

**PUBLISHED BY DIRECTION OF
THE CHIEF OF THE BUREAU OF NAVAL WEAPONS**

**For sale by the Superintendent of Documents, U.S. Government Printing Office
Washington, D.C., 20402 - Price \$3 (Buckram)**

TRAIL

contents**VOLUME 3****CHAPTER 1****INTRODUCTION TO VOLUME 3**

Synthesis of Systems	1-3
--------------------------------	-----

CHAPTER 2**WEAPONS SYSTEM REQUIREMENTS,
TASKS AND OPERATIONAL PHASES**

Functional Components of a Weapon System	6
Information and Material Systems	6-9
Information Inputs to System	8-9
Man-Machine Systems	10-13
Weapon System Tasks	
and Operational Phases	14
Target Detection Phase	15-17
Target Classification Phase	18
Target Location Phase	19-21
Weapon Selection Phase	21-23
Weapon Launching Phase	23-25
Weapon Direction Phase	26-30

CHAPTER 3**INTRODUCTION TO SENSORS
AND DETECTION SYSTEMS**

Electromagnetic Spectrum	32-35
Detection System Requirements	36-37
Methods of Detection and Tracking	38
Countermeasures	39

SECTION 1

INFRARED	41
Infrared Spectrum	42
Basic Physical Laws	43-45
Transmission Characteristics	46-47
Infrared Sensors and Detectors	48-50
Detector Characteristics	51
Infrared Seekers	52-53
Modulation	54
Infrared Receivers	54-56
Design Parameters	57
Military Applications	57

SECTION 2

RADAR	59
Types of Energy Transmission	60-64
Principles of Energy Transmission	65-67
Principles of Reflection, Refraction, and Diffraction	68-69
Antennas	70-75
Scanning	76-79
Fundamental Elements of Pulse Radar Systems	80-97
CW Radar Systems	98-104
Radar Countermeasures	105
Tactical Considerations	106

SECTION 3

SONAR	107
Physics of Underwater Sound	108-116
Transmission Characteristics	116-119
Sound Sources and Noise	120-122
Basic Sonar Systems	123
Transducers	124-128
Typical Searchlight Systems	129-132
Typical Scanning Systems	133-136
Comparison of Searchlight and Scanning Systems	136
Typical Listening System	137
Tactical Applications	137-138

SECTION 4

MECHANICAL AND MAGNETIC SENSORS	139
Inertial Sensors	140
Pressure Sensing Devices	140-141
Magnetic Sensors	141

CHAPTER 4**INTRODUCTION TO
COMPUTER FUNDAMENTALS**

Special and General	
Purpose Computers	144-145
Analog and Digital Computers	146

SECTION 1

ANALOG COMPUTERS	147
The Analog	147-151
Analog Devices	152-156
Road Mapping	157
Applications of Analog Computers	158
Solution of Equations	158-161
Simulation	162-163
Control	164
Limits and Stability	164-165
Scale Factoring	165-166
Programming	167-168
Summary	168

SECTION 2

DIGITAL COMPUTERS	169-171
Number Systems	172-174
Binary Arithmetic	174-177
Computer Codes	178-179
Boolean Algebra	180-183
Electronic Digital Devices	184-189
Memories	190-191
Arithmetic	192
Control	193-196
Input/Output	197-198
Operation of Digital Computers	199-205
CONCLUSION	205
Analog vs Digital Computers	205-207
Computer Applications to Weapon Systems	208-210

CHAPTER 5

INTRODUCTION TO COMMUNICATIONS

Communication System Requirements	212-213
Classification of	
Communications Equipment	213
Information Theory	214-218
Transmission Methods	219
Pulse Modulation Techniques	220
Methods of Communication	220-221
Modes of Intelligence Transmission	222-225
Man Machine Communications	226-227
Control and Displays	228-229
Introduction to Human Engineering	230-232

CHAPTER 6

INTRODUCTION TO SYSTEM DYNAMICS

SECTION 1

DEFINITION OF TERMS	235
Transfer Functions	236-239
Basic Parameters of Feedback Loops	240-241
Gain-Frequency Relationships	242-244
Order of Control	245-246

SECTION 2

DYNAMIC STABILITY	247
Sensor Noise	248
Tracking	248-250
Prediction	250-257
Weapon Station Motion	258
Prediction Cross Roll	259
Tracking Cross Roll	260
Weapon Line Cross Roll	261

SECTION 3

SYSTEM CONFIGURATION	263
Weapon Control Systems	264-265
Antenna Drive Tracking	266-267
Weapon Drive Tracking	267-268
Integrating Gyro	268
Computing from Weapon Line	269-270
Two Unit Tracking	270-272
Classifications	272-275

CHAPTER 7

WEAPONS SYSTEM DESIGN AND DEVELOPMENT

System Design Requirements	280-281
Determination of	
Operational Requirements	282-288
Fundamental Considerations	
in Weapon System Design	289-295
Methodologies - Tools	
of System Design	296-298
Systems Engineering	299-306
Chronological Phases in Naval	
Weapon System Development	306-307
The Complete Weapon System	307-312

abbreviated contents

VOLUME 1

CHAPTER 1

INTRODUCTION TO VOLUME 1

CHAPTER 2

INTRODUCTION TO
BASIC MECHANISMS OF COMPUTERS

CHAPTER 3

SYNCHROS

CHAPTER 4

INTRODUCTION TO SERVOS

CHAPTER 5

RADAR

CHAPTER 6

SONAR

CHAPTER 7

INTRODUCTION TO GYROS

VOLUME 2

CHAPTER 1

INTRODUCTION TO VOLUME 2

CHAPTER 2

INTRODUCTION TO WARHEADS

CHAPTER 3

INTRODUCTION TO
PROPULSION SYSTEMS

CHAPTER 4

INTRODUCTION TO
MISSILE FLIGHT PATHS

CHAPTER 5

INTRODUCTION TO VEHICLES

CHAPTER 6

INTRODUCTION TO
LAUNCHING SYSTEMS

CHAPTER 7

INTRODUCTION TO
WEAPON CONTROL SYSTEMS

APPENDIX A

REFERENCE FRAMES
AND COORDINATES

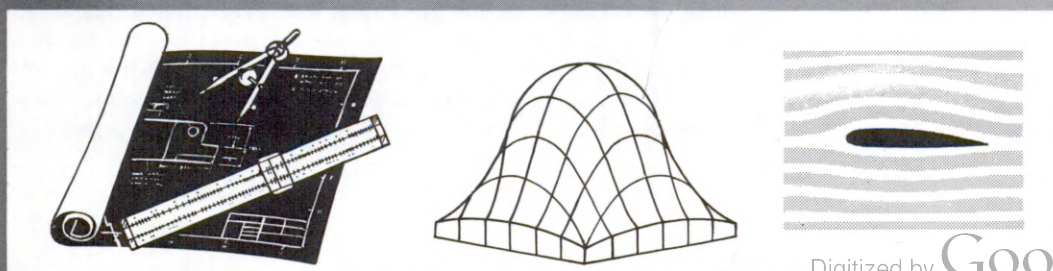
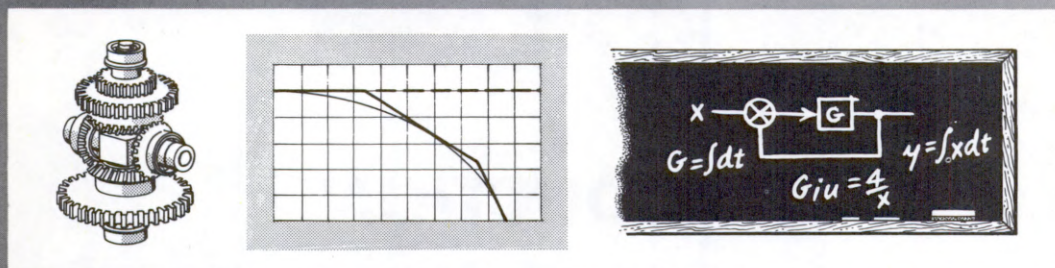
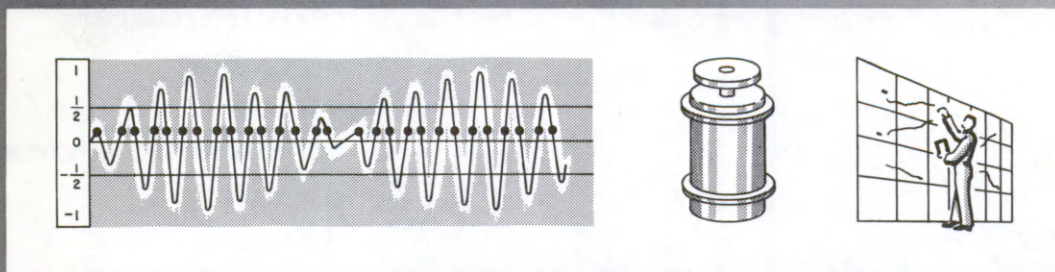
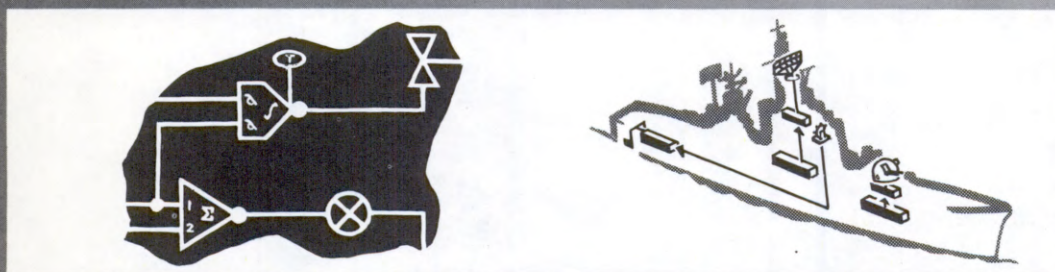
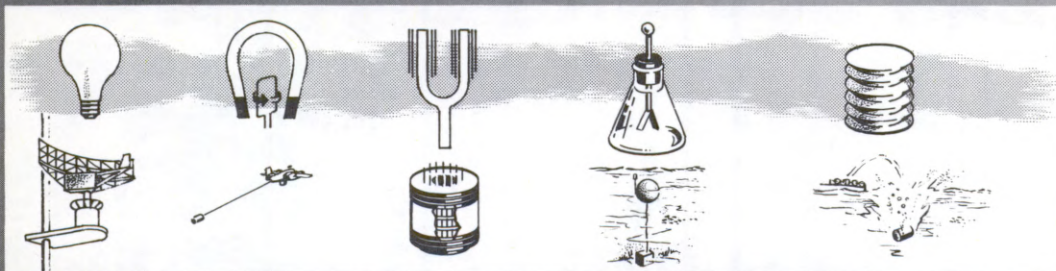
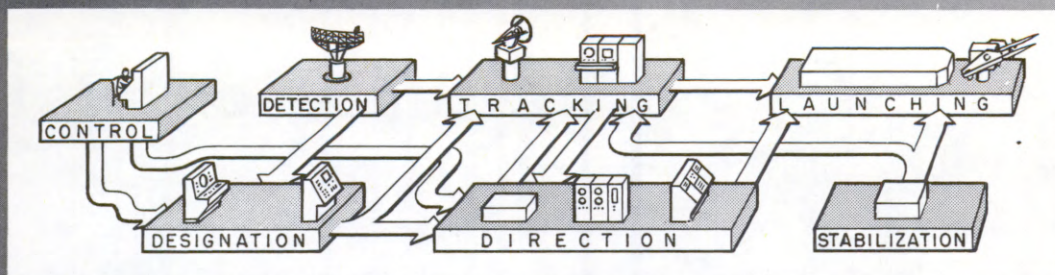
SYNTHESIS OF SYSTEMS



- requirements
- sensors
- computers
- communications
- dynamics
- design

INTRODUCTION

In the recent evolution of weapon technology, one of the most striking developments has been the transformation from the single weapon to the complex weapon system. Such a system is made up of a number of unique, specialized components which must be coordinated to achieve overall effectiveness.



SYNTHESIS OF SYSTEMS

synthesis . . . composition or combination of parts, elements, etc., so as to form a whole; also, the whole thus formed.

- ▶ **system requirements**
- ▶ **sensors**
- ▶ **computers**
- ▶ **communications**
- ▶ **system dynamics**
- ▶ **design and development**

system . . . an assemblage of objects united by some form of regular interaction or interdependence; an organic or organized whole.

system requirements

sensors

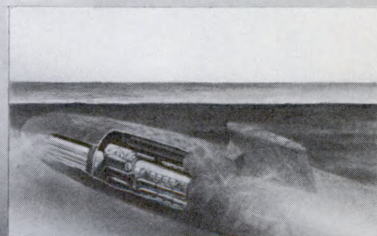
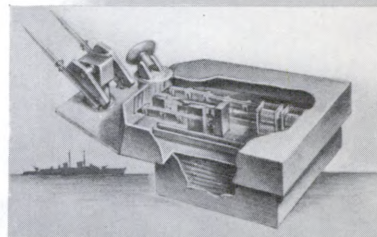
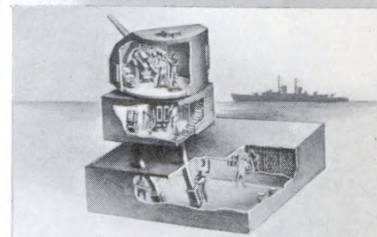
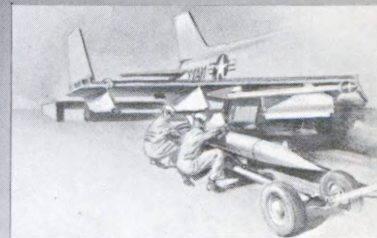
computers

communications

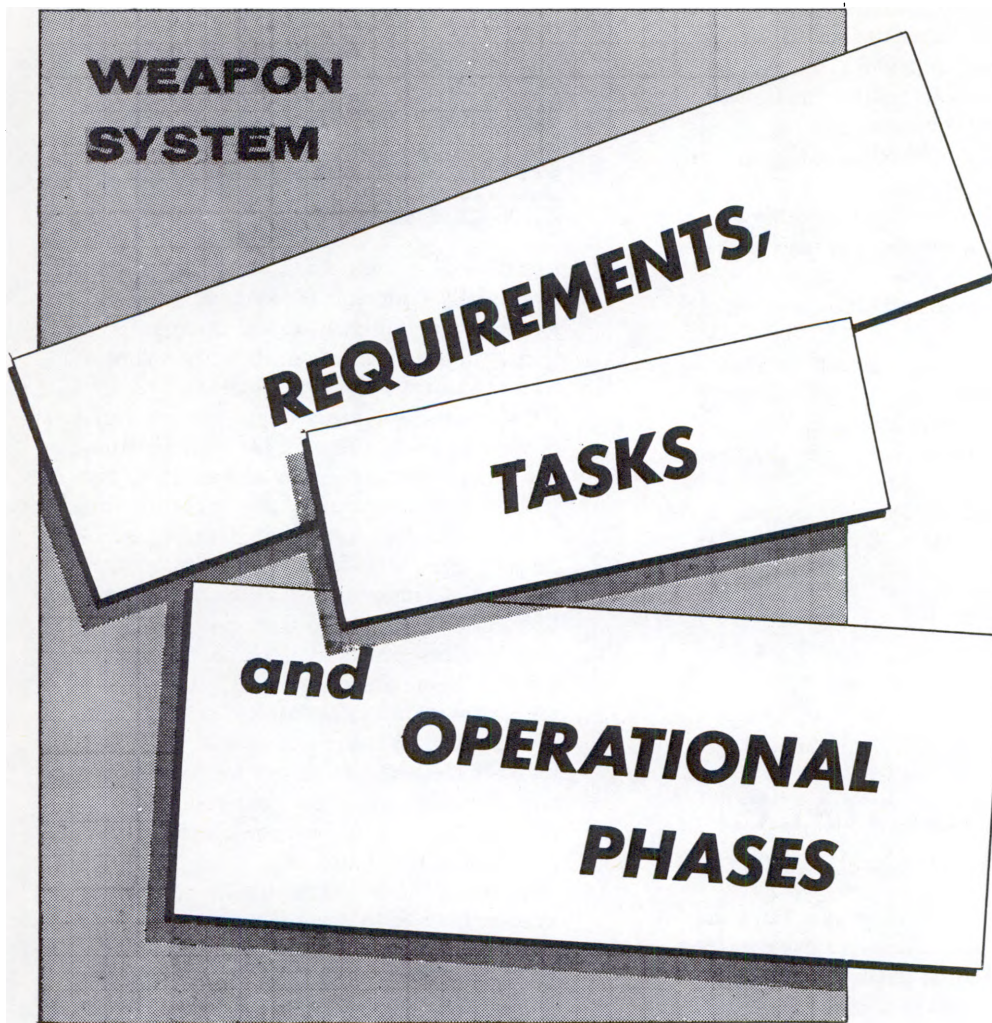
system dynamics

design and development

The basic components of a modern weapons system have been described in the first volume of this series. The second volume provides an introduction to the theoretical principles of advanced weapons and discusses some of the techniques whereby these principles are applied to weapon design. This volume is concerned with the solution of the problem of combining the various component elements.



The modern integrated weapons system must accomplish a series of tasks leading up to the achievement of effective damage to the enemy. Chapter 2 states the tasks to be performed by the weapons system. In succeeding chapters, some of the system elements are required to integrate the operational components of the system are discussed. Further, the problems of relating all of the operational components of the system are discussed in terms of the internal communications and the interrelated motion of components. In conclusion, a statement of some of the problems encountered in overall system design and engineering are discussed to provide an appreciation and evaluation of their bearing on the complete realization of the capabilities of the overall system. 3



A WEAPON may be defined as a device or means that may be employed against an enemy to achieve desired objectives or to defeat or deny similar objectives to the enemy. A WEAPON SYSTEM may be defined, therefore, as a collection of integrated components or subsystems which perform the interrelated functions necessary to render the desired effect on the enemy. This definition of a weapon system is intentionally very broad. It can include a man with a rifle, a guided missile system with its launching pad inside a submarine, a task force of submarines, or the striking power of an entire fleet. Regardless of the dimension of the target to be destroyed or the forces used to obtain the desired effect, the overall task of any weapon system remains the same: the delivery of a warhead or warheads to a target area or areas in a manner and at a time that insures the maximum probability of target kill.

FUNCTIONAL COMPONENTS OF A WEAPON SYSTEM

The definition of a weapon system states that it is a collection of components which operate together to render an effect on the enemy. The components of a weapon system, therefore, must perform certain related functions to achieve this effect.

The functional components of a weapon system are:

A **WARHEAD**, which inflicts the damage or other effect directly on the enemy target.

A **PROPULSION SYSTEM**, which provides the controlled release of stored energy necessary to move the warhead to the target.

A **VEHICLE**, which contains the warhead and other system components that travel to the target.

A vehicle and its contents comprise a **MISSILE**, which is any item of ordnance containing a warhead that is projected or propelled from a launching device towards a target.

A **LAUNCHING SYSTEM**, which puts the missile into the desired flight path in a safe and efficient manner.

A **WEAPON CONTROL SYSTEM**, which selects the proper flight path and controls the flight of the missile so that it follows the intended path to the target.

SIMPLE AND MULTIPLE TYPE SYSTEMS

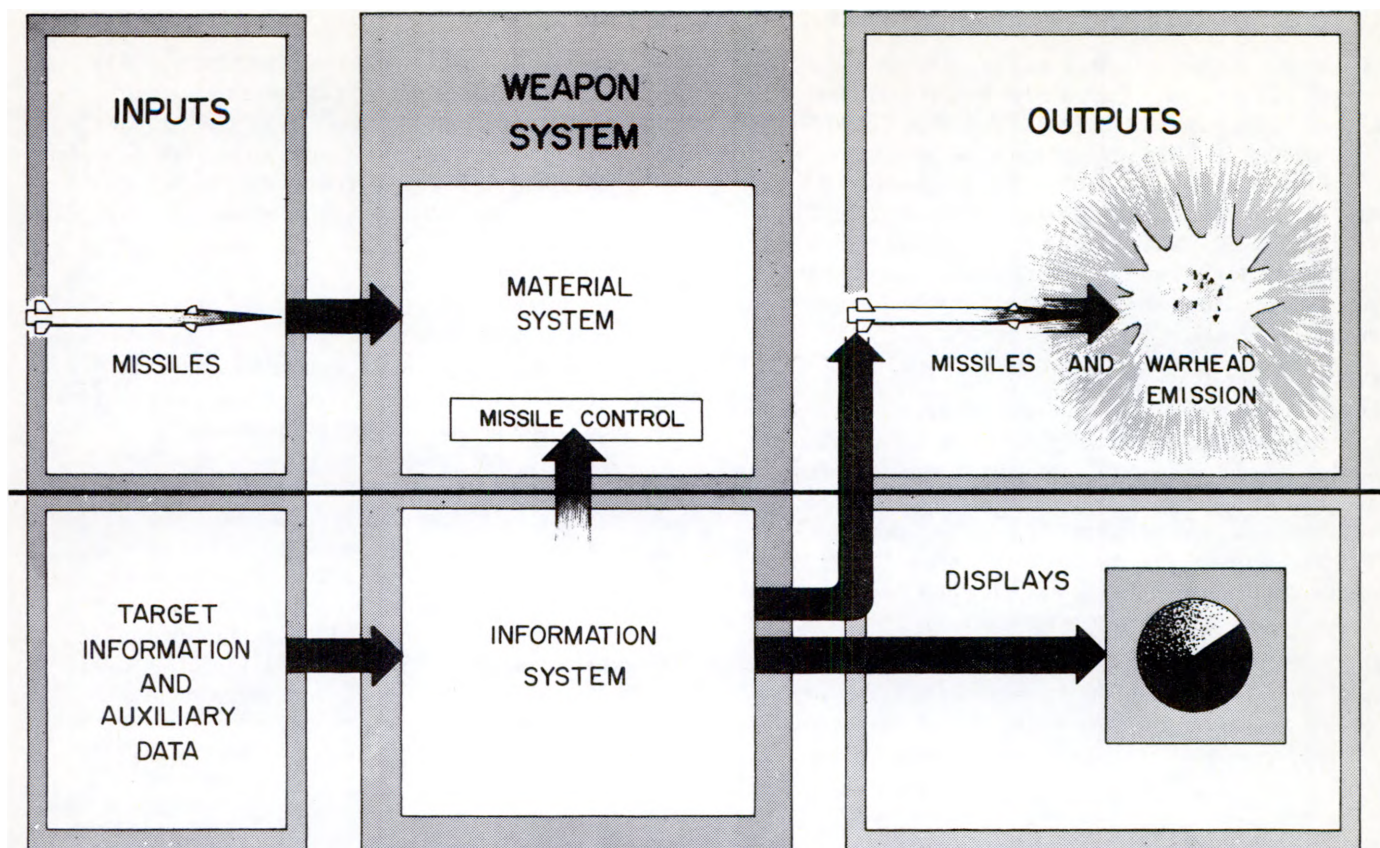
A weapons system may include only a single weapon or it may consist of several weapons; it may be designed to engage only one target at a time or several targets simultaneously. When multiple targets exist, the weapons system must be able to identify and select the targets in priority order or it must be able to engage several targets simultaneously, assigning each weapon its target to inflict maximum damage. To accomplish these tasks, a multiple weapons system must have a multiple target weapon control capability. The consideration of weapon potential and target threat is of prime importance in the proper choice of weapon or weapon combinations.

A man firing a rifle is an example of a simple weapon system. The warhead is the bullet; the launcher is the rifle itself; the charge in the cartridge serves as the propellant; and weapon control is obtained from the combination of the rifleman's judgment and the sighting mechanism on the weapon. In more complicated weapons systems, all of these basic components are systems within themselves. In a guided missile system, for example, the warhead is comprised of the fuze, the safety and arming mechanism, and the payload; the launching system consists of the launcher itself, storage space, and the transfer and loading equipment; propulsion is obtained from the rocket motor and the propellant, which is the rocket fuel. Finally, weapon control is obtained through the use of sensors and computers, which gather, process, transmit, and actuate the control device in a manner and to a degree necessary for target interception.

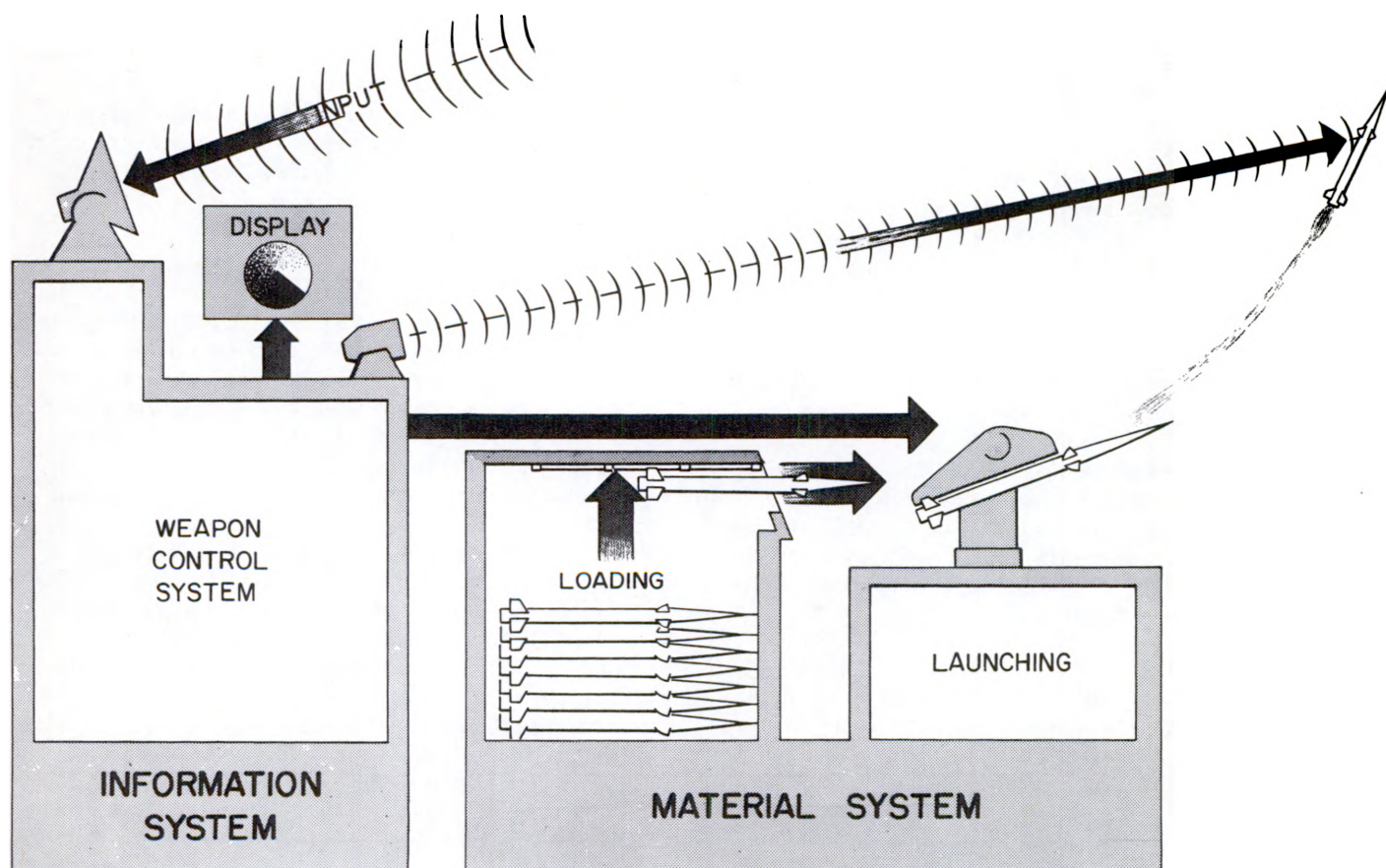
INFORMATION AND MATERIAL SYSTEMS

Regardless of their function, all systems are essentially information systems, and most systems are information and material systems. In the latter case, there is a flow both of information and of material within the system. For example, in an air-traffic control system, a flow of aircraft position and status information is necessary to control the flow of material (the aircraft) in the system. In a weapon system, a flow of target information is necessary to control the flow of material (missiles) to the target. The material inputs to a weapon system are missiles and associated propulsion units. The ultimate material output of a weapon system is the damage effector of the missile, warhead emission. The information inputs are target information and necessary ancillary data required to prepare the outputs, weapon control orders and information displays.

In general, the output of the information system is used to control the flow of material in the overall system. In a weapon system, the weapon control system is the information system. With its outputs, it controls the material flow of missiles, warheads, and warhead emission to the target. Its major output, the direction of the weapon line, controls the direction of flight of the missiles. The second output, information displays, is provided for the monitoring and decision-making human operators.



weapon system inputs and outputs



basic information processes

In any information system, there are certain basic processes which occur. For an information system to operate it must have information. Therefore, information must be GATHERED. To put the raw information into useful form, it must be PROCESSED. Inherent to the flow of information in any system is the need for DATA TRANSMISSION between necessary system components; and finally, for the system to efficiently accomplish its purpose, the information must be controlled or programmed correctly.

DATA GATHERING

Information gathering equipment includes sensors of all types, those which analyze motion (gyros, accelerometers, pitot logs), environment (barometers, altimeters, thermometers), and energy (infrared, radar, sonar). Energy sensors in weapon systems are used to detect and locate targets. This may be done actively, semiactively or passively, depending on the situation and the weapon. The data obtained must then be stored and processed.

DATA PROCESSING

Data processing equipment is required to store information and the appropriate mathematical operators needed for its particular computational role. Because

of the continuous nature and speed with which these computations must be made, computers have proved themselves indispensable to modern systems. On the basis of their computations, computers furnish the orders necessary both in sensor tracking and weapons direction. They also furnish continuous kill probability data.

DATA TRANSMISSION

To transmit data between machines and between man and machine, communication links must be established. Machines function in such a manner that they are dependent on each other. They are interconnected so that the operations or outputs of these functions in one will produce the corresponding functions in the other. An example of this dependence is a stabilization system employing a gyroscope in a servomechanism loop. The servo senses the stabilization error by means of an electric signal governed by a potentiometer, one end of which (the wiper arm) is connected to the gyro and the other connected to the servo. The potentiometer is mounted on the gyro case or gimbal, i.e., it is in the plane that is to be stabilized. Any deviation of this plane from the normal or gyro plane will cause a change in the output of the potentiometer proportional to the

INFORMATION INPUTS

The information system processes tactical information fed to it by its sensory information-gathering devices. The system must maintain a flow of information so that an overall picture of the disposition of friendly and enemy forces within range of the system intelligence is available at all times. The system must also permit rapid interpretation of each new bit of information, integration of new information with all other data, and immediate transmission of the new information with its significant implications to other components of the weapon system.

The inputs to the system may always be the same or of many types; they may occur periodically or very infrequently, they may be randomly distributed, or they may be countermeasure efforts attempting to confuse or destroy the effectiveness of the system.

The system can also be confronted with a low density information flow, a high density flow, as well as random, discontinuous flow, and must be designed to handle all three types. Information input types may be classified as single input or multiple input, and either can be at a fixed or random rate.

single inputs

If a weapon system utilizes one type sensor, which furnishes information at a fixed rate that is easily assimilated and analyzed by the prediction device, system components can be programmed to respond to the prediction outputs in the same manner at all times. This would be true even if the single input was received at an irregular or random rate, as long as the response time of the system is shorter than the rate of information flow.

multiple inputs

A system that comprises various sensor devices leads to a decision-making requirement for that system because there are then two or more possible types of input information. The more complex the type of input, the more complex the system must be to program this information flow in an optimum manner.

time distribution of inputs

If the response time of a system to any input is shorter than the shortest time between inputs, they can be handled as they come. If an input arrives so soon after the previous one that the system is still responding to the latter, then a queue or waiting line forms. If the average time interval between inputs is short compared with system response time, a queue of infinite length will form and the system will be overloaded. This situation may be handled by speeding up response or by providing more service channels. Most often, however, the inputs of interest to the system are distributed in time, i.e., the targets which are the inputs to a weapon system have a random time distribution. In this case, it is possible to allow system response to be longer than the shortest possible spacing (in which case a waiting line forms at some time) and to reduce the line's length at some subsequent time of longer spacing between inputs. Thus, a system may handle a waiting line in one of the following ways: more channels of service, faster response, or a buffer storage. The statistically distributed input is characteristic of a large class of systems.

degree of deviation.

In general, the communication links between machines are the energy couplings between them. Machines relay information to men through various measuring and energy-sensitive devices that can be read on a display or through a signal system. Oscilloscopes, meters, gages, and lights are examples of communication links employed between machines and man. This is only partial communication, however. The methods tell man what the machine is doing or sensing, but they do not provide a means to convey man's desires to the machine. Manual controls attached to appropriate machine circuitry or components provide the necessary return circuit by which man can command the machine.

Processed data includes both computations by computers and the judgments of the men involved. The main function of communication links is to provide a path by which the processed data can be used to achieve system control; good communication links are therefore an important part of system control.

DATA CONTROL

Besides the electromechanical and manual links needed to perform the operational phases of a weapons system (target tracking, weapon direction, launching, etc.) there must also be a control of the inputs to the system and the manner with which they are dealt. A means is needed to route or program the inputs. A single weapon system is usually unable to handle more than one target at a

time, although it can be kept informed of the location of the other targets by its search sensor. A multiple weapon system has the problem of assigning targets to make maximum use of its potential. In either case, when a weapon is assigned more than one target, a queue or waiting line forms whose length varies with the target input rate. In a multiple weapon system, targets assigned to a particular weapon or channel are not fixed, but may be changed or reassigned as the situation demands. The order in which targets are handled depends on their time of arrival (or detection) and their importance relative to both the attack and the overall mission.

RESPONSE

The response of data control equipment is dependent on the target input rate the system is expected to encounter. A problem involving a high rate of inputs requires fast data response. The ensuing weapon control phases must be correspondingly fast if the weapon system is to be at all effective against multiple targets or modern high-speed targets.

The output of the information system is used to control the flow of material (warheads) to the target. Note that the launcher drive system and the missile control system are considered part of the weapon control (information) system because they use processed information to control the flow of the missiles through the launching system to the target.

TO SYSTEM

countermeasures

Because countermeasures have become a standard part of military operations, weapon system designers are forced to consider the operation of their system in a countermeasure environment. In military weapon systems, countermeasure inputs are provided by a rational agent bent on the destruction of the system, thus making the system competitive in its choice of information. A distinguishing feature of competitive weapon systems is that they may operate in different ways on identical inputs and that the response to a particular input is not predictable.

The essence of competitive design is randomization of response. It should be obvious that in weapon systems such randomization is important. It is impossible to cover every aspect of military operations, regardless of probability, but it is equally impossible for the opponent in these operations to utilize every possible avenue of countermeasure. Therefore randomization of output is employed in many phases of weapon system operation. For instance, many weapon systems rely on radar as their sensing or information gathering device. Radar is easily jammed if its frequency is known. By having the radar system choose its transmission frequency at random, there is little chance of it being known before hand, even by the radar's operators. Similarly, since it is a monumental task to jam the entire radar spectrum, jamming apparatus may employ randomization of outputs. This is an example of the use of countermeasure and counter-countermeasure which form a vital area to all weapon system design.

counter-countermeasures

Just as the countermeasure designer must understand details of the operation of the various weapon systems that he must operate against, so the designer of counter-countermeasures must understand the various forms of countermeasures that his system is likely to encounter. To be effective, his system must be prepared for as many countermeasures as would be practical in a tactical situation.

A counter-countermeasure, therefore, is any means employed by a system to eliminate the effects of a countermeasure so that the system can still successfully perform the function it was intended for. This may be accomplished by increased training in target perception for the system operators, or by a highly technical electronic counter-countermeasure such as those employed against the more complicated radar countermeasures.

During WW II, the RAF hunted German U-boats with L-band radar. RAF had excellent results until the Germans installed L-band receivers in the submarines, thereby enabling them to detect RAF L-band radar signals at greater ranges than the British radar could detect an echo. This countermeasure was successful until the British realized the German tactic. They then developed a new radar of a higher frequency, and U-boat sinkings rose sharply. The German submarines on the surface, listening for L-band radar signals, were most vulnerable to RAF aircraft directed by the higher frequency radar. This is an example of weapon (L-band radar), countermeasure (L-band search receiver), and counter-countermeasure (new high frequency radar).

MAN-MACHINE SYSTEMS

Among the various target information factors a weapon control system must cope with in solving the weapon control problem are target type, quantity, threat, position, and velocity. Information concerning these various factors must be fed into the system. The system must then process these inputs and come up with an output or course of action. When the discerning intelligence involved is solely that of a man, as in the simple system of a rifleman, the problem is relatively simple. The man observes the situation and then makes his decision based on his own knowledge and previous experience. Such a system, while offering a great variety of outputs, is greatly limited by the physical capabilities of the man. As the target range and performance characteristics grow increasingly beyond the capabilities of man's senses, the use of machines to solve the fire control problem becomes increasingly necessary. The use of machines in weapons systems offers many advantages, especially in target detection and location and in the computations essential to target interception. Through the employment of such equipment as radar, sonar, and automatic computers, systems are able to function with tremendous speed and accuracy. Although machines have extended the destructive reach of weapons systems to points previously unthinkable, there is little probability of their completely eliminating the need for man. For all their advantages they are still machines designed by men to handle a particular situation in a particular way. No machine has as yet approached the ability of man to perceive and deal with an almost infinite variety of situations.

Before a machine is used in a system, it must first be established that it can perform the function more efficiently and more economically than man. Those operations whose accomplishments are impractical by machine are reserved for man. In almost all weapons systems man and machine work together communicating through a language of dials, meters, displays, and the like, to accomplish a particular mission.

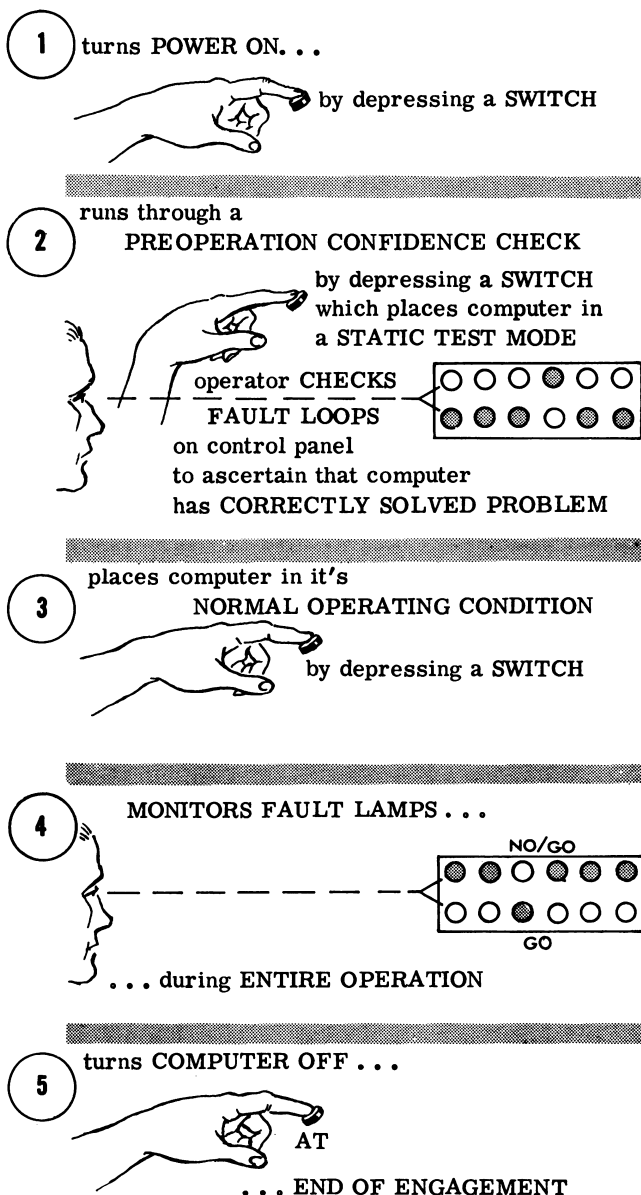
man as a component of man-machine loop

Although many weapon systems are described as fully automatic, they all use human operators in one way or another. In planning the design of a large system, a basic question which has to be answered by the designers is not whether man or automatic machines should be employed, but what functions should man be assigned in the system for optimum performance and efficiency.

In one type of system, primary control may be left in the hands of human operators who are assisted by data analysis, transmission, and display equipment. At the other extreme, systems may be designed which are almost fully automatic, and in which almost every function would be performed by machinery.

In these systems human operators are used for monitoring, checking, and maintaining the equipment. Between these two extremes lies a complete range of systems with varying degrees of human participation. Thus, a system could be designed so that its major work would be performed by semiautomatic machinery but in which the human operator would normally perform certain critical functions, for example, planning and decision making.

To illustrate the role of man in modern weapon systems, several types of equipments are discussed. In the case of the fire control computer used in a guided missile weapon system, the human operator performs the following functions:

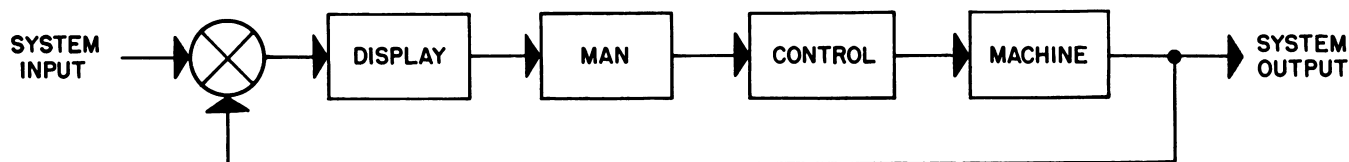


Of course, in addition to these functions, man also performs periodic routine and preventive maintenance tests as well as corrective maintenance when necessary. However, the extent of these duties may also vary since solid state technology has reduced the preventive and routine maintenance requirements and also because of advances in built-in fault locators.

The role of man in a tactical display system is quite different than the one of the computer operator described above. This system may require several operators whose function is to track incoming targets on various scopes and also an officer whose function is to assign priorities

to the various targets based on displayed data, such as size, threat, position, etc. In this case, the control of the entire weapon complex is placed directly with man. His decisions are aided by the machine-made detections and calculations presented to him, but the decision rests with him alone. Regardless of how complex a weapons system is, there exists a necessary correlation between man and machine, a man-machine loop, in which man plays the guiding role.

A typical man-machine loop is illustrated here for purposes of discussion.



The loop consists of an error differential which subtracts the system output from the input to produce an error signal. This signal is then fed to a display which usually portrays continuously changing information. The operator senses this information and performs his functions accordingly. Thus, the output of the machine (the system output) is regulated by the control settings of the human operator.

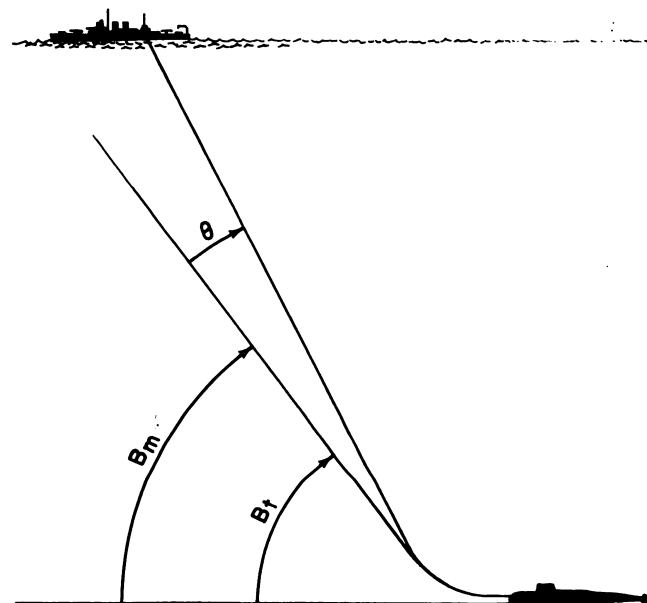
A man driving an automobile is an example of a closed-loop man-machine system. The input to the system consists of road and traffic conditions which are displayed to the operator by means of the automobile's windshield. Based on the information displayed, the operator uses the controls (brakes, clutch, steering wheel, etc.) to make necessary manual adjustments, thus safely maneuvering his vehicle to its destination. It is a closed-loop system since its output is constantly fed back as input through the windshield. Thus the operator is continually evaluating the output of his machine.

A military example of a man-machine loop is a wire-guided torpedo system. In this system, the input is target relative bearing, the output is torpedo relative bearing. The object of the operator is to guide the torpedo to intercept by maintaining zero error between the target and torpedo relative bearing.

The torpedo relative bearing is subtracted from the target relative bearing to obtain the error in relative bearing. This error is displayed on a dial which is monitored by the operator who rotates a handcrank to null the error and to maintain the null. The output of the control, the required torpedo orders, is fed to the machine (in this case the torpedo) which accepts the required torpedo orders and corrects the torpedo relative bearing. The new torpedo relative bearing is fed back to the error differentiator.

Even in fully automatic systems, the human operator is necessary because of his flexibility and his ability to monitor the machine. Although the machine may perform all the computations and even make decisions, the human operator's flexibility is put to use by providing him with an override capability.

The machine is relatively inflexible because of the size of its memory and the limitations of its program. It cannot make decisions outside of its experience, and therefore requires a human operator to make them or to override a wrong decision that the machine might have made. In many cases, the operator is required to control the operation of the machine by feeding it inputs which guide the machine to the proper solution. These control inputs are considered decisions. From this discussion, it can be seen that the operator's function may vary from one of monitoring to one of complete control, with all of the possible incremental differences throughout this range.



B_t = TARGET BEARING
 B_m = TORPEDO BEARING
 θ = ERROR

human engineering

Human engineering is the study of the physical and mental limitations of the operator, that is, his endurance, speed, and accuracy, as well as the psychological aspects of his tasks. As man-machine systems become more complex, the tasks assigned to man often become more complex, also. Many aircraft accidents that had been originally considered a result of pilot error were found upon further investigation to be the result of poor human engineering. That is, the tasks assigned to the pilot proved to be beyond his human capabilities.

In designing a machine to be operated by man, consideration must be given to the senses of the operator upon whom the successful performance of the machine rests. When controls are operated manually and the operator depends upon his sense of touch to distinguish between them, each control must be given a distinguishing shape and all controls must be within easy reach of the operator.

If there are readings or observations to be taken by the operator, the dials or displays from which they are taken should be designed in such a manner so that both their appearance and position afford maximum legibility to the operator. Also, visual alarms such as warning lights should be located where they are most apt to be seen without injuring overall machine effectiveness. Audible alarms such as buzzers or bells must be distinguishable from the noises inherent to the equipment. The study of these matters in designing equipment is essential for optimum man-machine efficiency.

man and machine bandwidths

Because man is often employed as an active element in a weapon system, his response to both steady state frequency inputs and transient inputs is of interest. Man is a nonlinear system. However, for low frequency inputs, and for operations within his performance ability,

capabilities and limitation

Responsibilities for the various functions of a modern weapon system are assigned after determining what man can do better than machines and vice versa. For instance, if inputs to the system consist of typewritten data which vary in size, form, or arrangement, and the first function is to read these inputs and transform them into code form for computation, there is little question who should have this responsibility. As yet, there is no known machine which can perform this function.

The following list gives some examples of the relative capabilities and limitations of man and machine. Lack of training, ineptness, inability, fatigue, boredom, and environmental conditions, all effect the performance of man. The failure of man to perform his operations properly and to make correct decisions will effect the probability that the weapon system will succeed in destroying the target. Thus, the effectiveness of the weapon system is directly affected by human decisions.

MAN	MACHINE
Able to handle low-occurrence probability inputs	Virtually impossible to program for all possible events
Able to organize many small bits of information into meaningful and related wholes	Organization programming is difficult because of the various ways of organizing different types of information
Able to achieve satisfactory results by alternative methods if primary means fail or are damaged	Alternative modes of operation limited; may break down completely when partial injury or damage occurs; not able to regenerate or heal
Able to handle only a small amount of information per unit of time	Information handling capacity can be made almost as large as required
Able to operate efficiently only over relatively short time periods because of fatigue and boredom	Efficiency normally decreases only over long periods of time
Able to change programming easily and often; large variety of programs available	Program changes and variety can be achieved only at a great cost
Not able to make accurate computations swiftly	Excellent and very rapid computers

his output may be considered linear. The response characteristics between men also differ, but there is an area of operation wherein the human characteristics are sufficiently similar. This is the area that the designer of a system must employ.

For illustrative purposes, assume the following example: A man is assigned the problem of following a steady state frequency input, such as a spot of light moving with a sinusoidal motion in a single plane. If the mechanism he is employing to track the single spot is within his physical capabilities of handling, he will behave as a linear power amplifier to a frequency of one or more radians per second. When the frequency is increased so that he no longer can track the moving spot, his response becomes completely unpredictable. If high frequency noise is superimposed upon the fundamental signal frequency, the bandwidth of his linear operation is reduced. This would imply that noise should be removed

from any presentation to the man, and that he may best be employed as an amplifier at low frequencies.

In the case where a man is optically tracking a target which is moving at constant velocity, the angular rate of the target motion, as seen by the man, is not constant, but changes gradually. If the tracking device employs a rate memory element, the input to the man becomes the difference between the angular rate of his tracking device and the observed angular rate of the target. This is called rate aiding or aided tracking. Under the condition of rate aiding, the bandwidth of the aiding device and the man combined becomes considerably greater than if the inputs to the man were the full target motion. The bandwidth of the man may be reduced by his environment. If, in the process of his operation, he is subjected to noise, heat, shocks, or other disturbing influences, his bandwidth is reduced.

With the technology now available, it is possible to replace the human decision by the computed decision. A computer is faster, and if the inputs to it were correct, it would make fewer errors of judgment than man. Considering the present status of computer design, man is still required for the following reasons:

- 1) He requires less space, at least insofar as the computational mechanism is concerned. Although the mechanized computer has the principal advantage of speed over man, the human brain in mechanized form (if this were presently possible) would be tremendous in size, using presently available techniques.
- 2) He is more versatile in the event of an emergency. To keep computers to a reasonable size, they are highly

specialized in the memory elements and problem solutions. The computer is adapted or set up for the solution of a particular problem or problems and cannot be changed.

- 3) He is required to maintain the computer in operation, and the training for this function is usually more specialized and time consuming than the training required for the performance of the operational functions connected with the weapon system.

One element that will make it necessary that man be displaced from his present role of decision making is that of time. When the time available becomes too small to permit the use of man as part of the system (because of increased target speed or other reasons), then he must be replaced by some mechanized device.

allocation of tasks

The allocation of tasks to man and machine in a weapons system is necessarily a function of the relative capabilities of each. It follows that the state of the machine art is a major factor in the definition of a specific man-machine loop.

The following functions are usually reserved for man:

- 1) Qualitative memory storage of doctrine, procedure, mission, objectives, etc.
- 2) Improvisation to handle events of low probability (Since it is virtually impossible for machines to handle all situations, they are necessarily geared for the higher probability events.)
- 3) Application of judgment, particularly in the selection of alternate courses of action or in decisions based on original information
- 4) The use of perception in recognizing the significance of complex data (i.e., recognition of enemy tactics or intentions)
- 5) Evaluation of data from multiple sources, particularly in case of conflicting data

The following functions, where consistent with the state of the art, are assigned to the machine:

- 1) Quantitative memory for the storage of computational programs and detailed data such as problem and ballistic information
- 2) Repetitious or routine operations which are tedious and tiring for man
- 3) Power assistance for remote and difficult operations
- 4) Rapid and accurate sequencing of operations whose timing and accuracy are critical
- 5) Rapid and accurate solution of computations

Man, of course, has his place in the solution of the weapon control problem. In the design of a weapon system, each phase is examined and after the alternatives are weighed the various operations to be performed may be designated to either the man or the machine. To insure the maximum system "up-time", the functions to be performed by machine and man often supplement one another. An examination of the operational phases in accomplishing the weapon system tasks provides an excellent application of the man-machine concept.

WEAPON SYSTEM TASKS

The overall task of any weapon system may be generally described as the delivery of a warhead or payload to a target area to insure the maximum kill potential of the system. A basic analysis of this task points up two major divisions or subtasks which must be accomplished to attain the desired kill probability.

TASK	TASK
TARGET DEFINITION	WEAPON DELIVERY
OPERATIONAL PHASES	
TARGET DETECTION	WEAPON SELECTION
TARGET CLASSIFICATION	WEAPON LAUNCHING
TARGET LOCATION	WEAPON DIRECTION

The **TARGET DETECTION PHASE** involves the surveillance of a known energy field and the detection or discrimination of any anomalies that appear within that field. For example, the energy field may be the field of electromagnetic radiation from the sun, or Earth's magnetic field, and detection would be achieved by visual or magnetic sensors. The energy field can also be man-made, by light, radar, or sound energy sources being located either remotely or within the weapon system. In either case, suitable energy **SENSORS** are needed both for surveillance and detection of anomalies in the energy field.

The **TARGET CLASSIFICATION PHASE** consists of classification of the detected anomaly and identification of its source. Classification involves an analysis, perception, and definition of the nature of the detected anomaly. Identification of the source determines its friendly or unfriendly character. For example, a sonar operator performs a surveillance of an active acoustic field. When he sees a blip on the sonar display or hears an echo, he analyzes these signals to determine their source. By comparing the sound signals received with the sonic signatures of known sources, he can classify the detected anomaly. If the classification process indicates that the anomaly was caused by a submarine, the operator identifies the character (friendly or unfriendly) of the submarine by whatever means are available.

The **TARGET LOCATION PHASE** consists of locating the target relative to the weapon station with sufficient accuracy for effective weapon employment, and the maintenance of sensor contact with the target for protracted periods. This phase requires the employment of sensors to maintain contact with the target and to sense information about its location and motion.

The **WEAPON SELECTION PHASE** involves selection of the optimum type of weapon, consistent with the mission of the weapon station and its capabilities, and the design-

ation of a specific weapon system to destroy the target. For example, the captain of a destroyer may have depth charges and antisubmarine torpedoes available for employment against a submarine target. He will select the type of weapon (from those actually available and ready for use) which appears to have the highest kill probability under the prevailing conditions, and he will designate the related weapon system to destroy the target. In a guided missile cruiser, on detecting a threatening air target, the weapon control officer will select either a beam riding or homing missile and designate a specific guided missile control system to the target with that type of missile.

The **WEAPON LAUNCHING PHASE** is concerned with the safe and efficient launching of a missile into the desired flight path. It involves also the assessment of target damage and the preparation for reattack.

The **WEAPON DIRECTION PHASE** consists of acquisition of the target by the designated weapon control system and the generation of the necessary weapon control orders to intercept the target. Acquisition requires that a sensor within the weapon control system acquire and lock-on the target to gather information pertaining to target position and motion. This information is processed in the weapon control system to generate weapon control orders. The weapon control orders define the orientation of the missile velocity vector needed to destroy the target. This information is sent to the launching system for proper orientation of the launcher and the missile it contains.

In this phase, a target detecting sensor with an appropriate drive system is needed to continuously gather information about the target, and a data processing system (computing system) is required to transform the sensed target data into useful weapon control orders. In addition, a drive system is required for orientation of the launcher in response to weapon control orders.

target detection phase

Target detection requires continuous, long term methodical surveillance of a known energy field, either natural, locally propagated or remotely propagated, and the ability to discriminate an anomaly within that field.

means of detection

The general methods employed in target detection may be classified in three groups: passive, active and semi-active. Passive detection senses radiant energy emitted from a target. Detection is considered active if the transmitter emitting the energy and the receiver sensing the energy reflected from a target are both located on the same vehicle. In semiactive systems, the receiver senses the energy reflected from a target being illuminated by a transmitter at another location.

PASSIVE DETECTION

Of the three basic methods of target detection, passive detection requires the least equipment. This detection method uses emitted energy from the target as a source of information. Since the target is the source of energy, no transmitter is necessary. The energy may be acoustic, magnetic, thermal, or electromagnetic radiation. Sound as a source of information has its main application in underwater probing and is used in sonar systems. Passive detection by sonar is essentially a matter of listening.

The success of detection by listening is primarily dependent on the ability of the operator to evaluate the sounds. It may be simply a matter of hearing ability or it might be the ability to distinguish various sound wave shapes or signatures on a display. This latter method is used particularly in ultrasonic listening since these sounds are out of the audible range. Ultrasonic sounds are also converted to the audible range by heterodyning in the receiver. Listening often gives bearings quite accurately, but provides little or no information on the range, except in specialized equipment.

Listening is used chiefly by submarines, since surface vessels at high speed produce considerable noise which interferes with the detection of the sounds of other ships, especially the low frequency sounds of submarines. On the other hand, this difference in noise output enables a submarine to detect the presence of a surface vessel rather easily.

Radio and radar are especially useful in missile design and in combatting aircraft. A commonly employed passive detector using this energy is the radio direction finder. This same general technique can be used by detection systems in guided missiles. However, detectors of this type depend upon the target's radiation of radio or radar energy. If the target maintains radio and radar silence during an attack, this form of detection is useless. For this reason, most practical systems consider only those sources of energy which are inherent to or are uncontrollable by the target. Heat and light are, therefore, the sources of energy of the greatest interest. The infrared (IR) portion of the electromagnetic spectrum is of particular interest because a major portion of the radiation emitted by targets (jet engines, for example) lies in that region.

Several means are available for detecting the presence of IR energy. Those most commonly used devices are either heat or light sensitive. Thermal detectors include thermocouples and bolometers (both rely on electrical response to heat energy, bolometers being more sensitive). When a thermal detector is exposed to IR radiation, its temperature rise or response is proportional to the net exchange of radiant energy between object and detector.

Those devices sensitive to light are called photoconductive cells. Unlike the thermocouple or bolometer, the photoconductive cell does not depend on heating for its response. Response is a function of the intensity and wavelength of the incident light.

However, the electromagnetic spectrum is not the only source of energy useful for passive target detection. A target may emit other types of energy fields, such as magnetostatic and electrostatic, which can be used to detect its presence.

Visual detection devices, although not as common as they were once, are still very much in use and are for the most part considered passive. The eye, telescope, optics, television, etc., are all important means of target detection.

ACTIVE DETECTION

Active detection systems require both a transmitter and a receiver of energy on the vehicle. Searchlights, radar, and sonar are all systems which can be considered active. Searchlights are still used in visual detection. On occasion they have even been used to locate underwater targets. The use of searchlights gives a ship an emergency alternative; however, underwater targets are normally detected by acoustic and magnetic detection equipment.

Active sonar consists essentially in emitting sound energy from a transducer and detecting a reflection of this energy which indicates a discontinuity in the region being surveyed. The energy reflections may be sensed aurally through the earphones of the operator or visually by means of a display on a cathode-ray tube. The active sonar system is commonly found on surface ships where noise conditions make passive systems less effective at high speeds.

Radar systems all use the echo principle, the detection of a target by reflecting energy off their surfaces. The amount of energy reflected by the target depends on the reflecting surface material, the cross-sectional area, and the aspect angle. Three different methods have been used in radar: frequency modulation, continuous wave (or frequency shift), and pulsed radar. Pulsed radar is the most commonly used type of radar in ship-board installations.

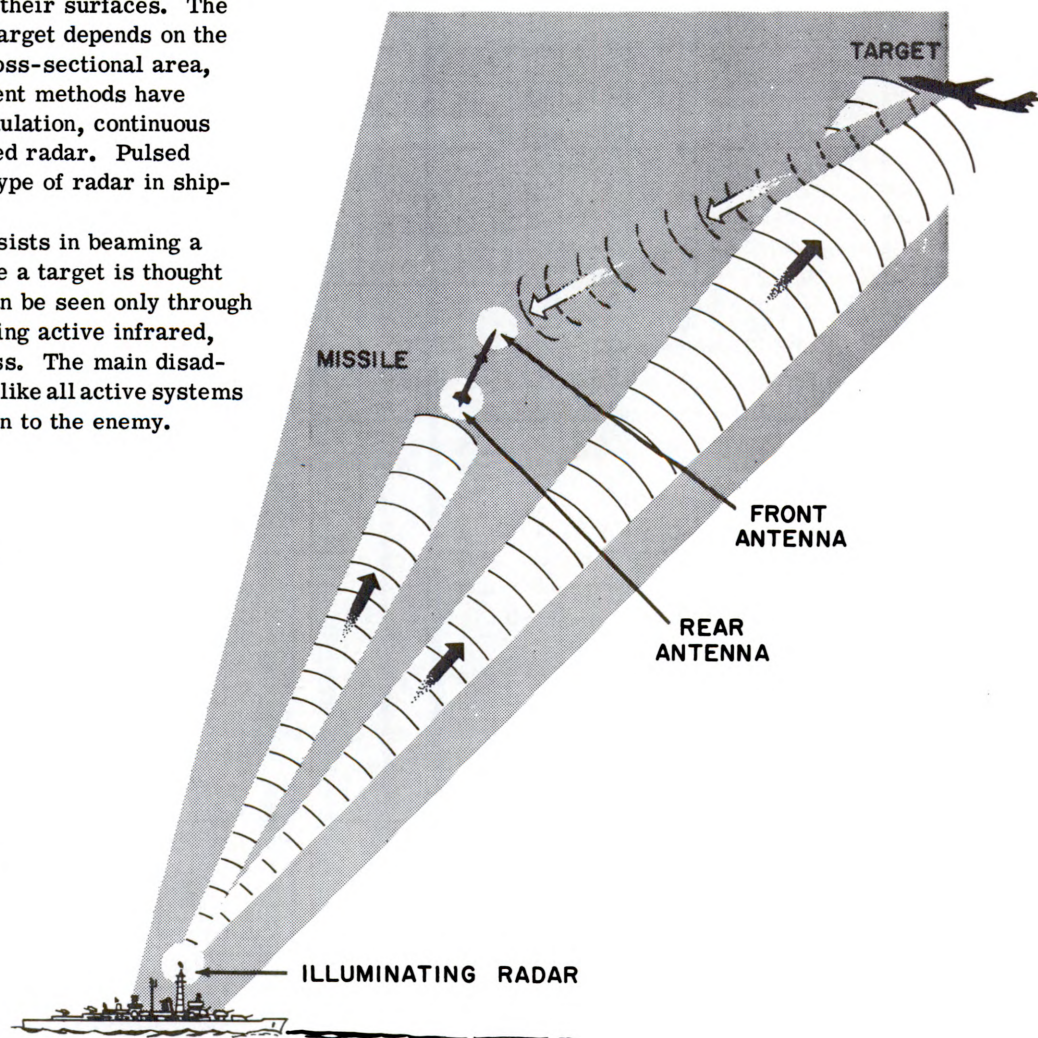
Active detection using infrared consists in beaming a source of infrared on an area where a target is thought to exist. The energy reflections can be seen only through an infrared sensitive device. By using active infrared, targets can be seen in total darkness. The main disadvantage of this active system is that like all active systems it can betray the observer's location to the enemy.

SEMIACTIVE DETECTION

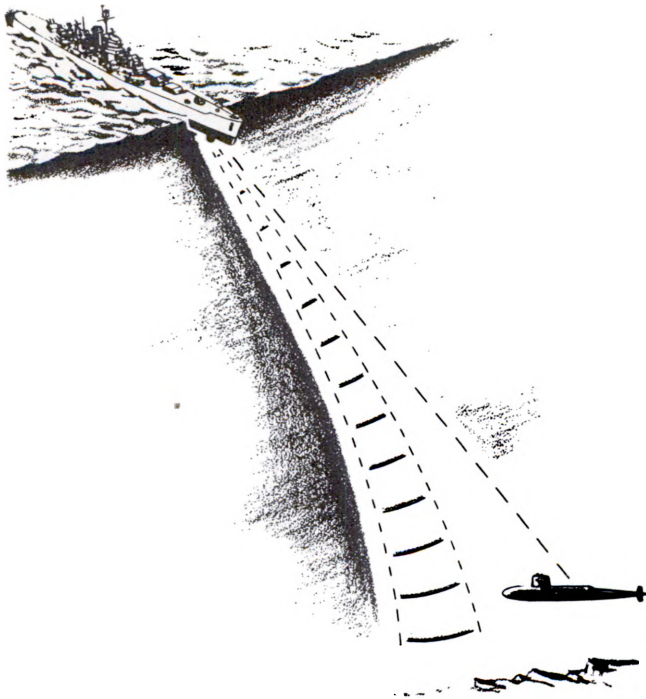
To gain the advantage of a high powered, long range sensor, which the missile cannot carry, weapon systems may employ semiactive detection systems in which the energy transmitter is located external to the missile, on the launching vehicle, while the receiver sensing this energy's reflections off the target is located in the missile. Thus, while not radiating electromagnetic energy, a missile can home in on the target-reflected energy transmitter from an external source.

attenuation

In all phases and methods of target detection there are certain problems in energy transmission and equipment limitations that must be dealt with. All energy wave propagations are subjected to varying degrees of attenuation, depending on the nature of conditions in the transmission medium at the time. A target may be within normal detection range under average conditions but, because of some environmental condition, the reflected energy is sometimes indistinguishable from noise and the target may proceed undetected. Environmental conditions in the transmission medium can cause the energy beam to be reflected, refracted, and absorbed. As a result, the energy beam may be bent and curve away from a target location. In other cases, it may be trapped or ducted, increasing the energy intensity in some areas and eliminating it in others. The net effect is generally to reduce the probability of target detection.



semi-active detection system



There is also the problem of ghosts which results from more than one reflection path, as in the case of a radar picking up energy reflections directly from an aircraft and also those bounced from the aircraft to the water. If the ghost image is strong enough, it is very difficult to distinguish between it and the actual target and the system is likely to go after the ghost. A skilled operator is effective in preventing this type of system error.

Attenuation may also be caused by climatic conditions such as rain, snow, or fog which absorb and disperse energy waves. This is especially true with infrared detection, where a heavy precipitation blinds the infrared detector by completely hiding the infrared radiations of the target. Similarly, radar and sonar are blinded by physical obstacles such as mountains, heavy vegetation, etc.

Blind areas in the sensor device may be the product also of equipment limitations in elevation and traverse or in sensitivity. Radar detection systems are often defeated by aircraft flying beneath the radar scanning volume. Because of these blind areas, visual means of detection are still vitally important.

target characteristics

In addition to the problems inherent to target detection, it is important that some consideration be given to the target characteristics and tactical environments with which a weapon system may be required to operate.

The target may be single or multiple. At long ranges, the sensing equipment may not be able to distinguish between one target and a group of several targets. As

the range shortens, a point is reached at which the sensing system can distinguish single targets. This quality is known as target discrimination. Until this point is reached, the weapon system is being directed at the electromagnetic center of gravity of the formation with consequent likelihood of passing between targets. If we choose a sensing system of good target discrimination, the enemy will find it difficult to defeat the system, even by employing optimum spacing of targets in his formation.

In combatting the multiple target problem a system may employ multiple sensors or a system of channels. In using multiple sensors each sensor can concentrate on a particular target, or it may handle a segment of the target area and relay target information to the rest of the system. A single sensor handling many targets must assign target channels either in order of arrival or in order of threat. A queue will then form because the system dispenses with the targets only one at a time. The decision as to target importance is usually reserved for the operator.

countermeasures

The enemy may also employ chaff or decoy devices calculated to throw off our detection system. Again, this adds undesirable weight, an especially critical factor in an aircraft. However, we can postpone effective countermeasures by choosing a detection system which is relatively invulnerable. By parallel development of improved or alternate detection systems, we can be ready to continue operations when and if enemy countermeasures do become effective. The importance of exercising a high degree of security over information which the enemy could use to design effective countermeasures is evident.

The enemy may also employ electronic countermeasures to defeat the detection system. To do this effectively he must ascertain our system characteristics and develop, test, and produce jamming or deception equipment that can be installed in his ship, aircraft, etc., with tolerable increase in weight and decrease in performance.

man-machine link

There is a necessary man-machine correlation involved in successful target detection. Information is gathered through sensors, processed and transmitted to highly trained operators by means of appropriate communications and displays. The success of the system relies not only on the speed and sensitivity of the machine to handle target data (sensors, computers, communications links, displays) but also on the ability of the men involved to interpret and use this data. Among the many questions posed which are usually reserved for men are: What is the nature of the target indicated? It is an actual target or a decoy? It is one target or many? Are there two targets or is one real and one a ghost? Which is the real target? After supposed intercept, was the target missed, destroyed, or just slightly disabled? In their present state of development, machines cannot be considered to have decision-making ability; this right is still reserved for man.

TARGET CLASSIFICATION PHASE

In its broader sense, the target classification phase encompasses not only the classification of the target, but also the identification of the friendly or enemy characteristic of the source. Based on this information,



a decision concerning the proper course of action (attack, evade, ignore, etc.) consistent with the vehicle mission and tactical situation can be made by the commanding officer.

information requirements

The target classification phase imposes three general information requirements. (1) Data which must be sensed is the detected anomaly and the communication from your own force. (2) Data which must be stored includes the frequency spectrum of known targets, typical target tactical behavior patterns, friendly forces identification codes, operation plans, operational communications, and intelligence reports. (3) Data processing capability required consists of spectrum analysis of anomalies detected and comparison of the analysis with stored information of known target characteristics. In most weapons systems, the information gathering or sensing requirements are fulfilled by specialized machines such as radar, sonar, and communications equipments. Man is usually used for data storage and spectral analysis, but machines may soon be able to do this job. The most promising approach seems to be a machine adaption of the method now used by man, i.e., correlation of the power spectrum of the signal with the spectrum of known sources.

target analysis

During the target detection phase, frequent distortions or irregularities in the generated (or ambient) energy field will be detected by the target sensor. Each of these anomalies must be processed as a possible target. The first stage in the processing is a search for any distinguishing characteristics and an attempt to identify them.

Detected signals are characterized by distinctive signatures (or frequency distributions) when subjected to a frequency spectrum analysis. Identifying characteristics peculiar to a target of interest can be obtained therefore by observing selected valid targets. The observed data can then be stored in a memory unit for later comparison or correlation with the characteristics discernable in a sensed signal. Consequently, the signal processor must consist of a memory unit capable of storing characteristic target spectrum and a means of recognizing these characteristics in a sensed signal. classification. External verification, while not essential to target classification, can be very useful in defining a target whose nature is doubtful.

target behavior patterns

Signal behavior is an important factor in the classification of detected anomalies. Signal behavior relates to the signal source activity (i.e., motion and rate of

motion) apparent in the signal. A signal source having an extremely erratic behavior (erratic bearing rate, for example) does not indicate as credible a target having a relatively smooth bearing rate. Furthermore, the degree to which any apparent motion is purposeful over a long period of time is an important clue to classification.

geographic targets

Additional information factors useful in target classification are the geographic situation and intelligence data. These are especially helpful in defining stationary targets such as mountains, reefs, shoals, and navigational buoys.

IFF

However, for moving targets, identification which relies on known or expected location data is less dependable and less useful than a communicated signal. To afford protection to friendly forces, IFF (Identification, Friend or Foe) electronic equipment is used in conjunction with sensor systems. Through the use of IFF equipment, craft can be identified as friendly without resorting to verbal communications or visual identification. In the IFF base, a transmitter emits a signal on a particular frequency. The signal is received by the craft in question and in turn transmits a coded signal back to the base station. If the base station receiver is set to the particular coded pulse chain that is returned, it automatically signals the operators that the craft is friendly. If no signal is returned or the code is incorrect, the craft is assumed to be unfriendly.

Initial target classification must proceed from a survey and evaluation of the preceding information factors. Further confirmation or contradiction of original classification can be established on the target location and motion analysis. Additional information factors govern the command decision as to the course of action to be taken when the anomaly has been classified as a valid target. Data made available by prior intelligence should indicate the relative capabilities of the target. Command must balance this knowledge against his vehicle's capabilities, its mission, and its condition of readiness and vulnerability. Finally, as target analysis becomes available, prior action decision can be confirmed or modified by apparent enemy intentions.

TARGET LOCATION PHASE

The target location phase is defined as the location of the target relative to the weapon launching vehicle. Information gathered during the target location phase must be accurate enough to insure effective attack and the maintenance of sensor contact with the target for a protracted period (usually until after target interception has been confirmed).

information requirements

To accomplish the target location phase, the following information factors must be determined:

- 1) Target bearing
- 2) Target range
- 3) Target elevation
- 4) Sensor platform orientation
- 5) Energy path prediction
- 6) Own location
- 7) Parallax compensation
- 8) Target motion

Target bearing, range, and elevation fix the target position in space with respect to the weapon system. If the target is geographically fixed, these factors can be computed from the location of the sensing system. If the target is moving, target location must be established by data sensors carried by either the delivery vehicle or a consort. In either case, the performance characteristics, such as accuracy of the required inputs concerning the target, or speed of the handling of these inputs, are largely a function of the weapon capabilities and the target rate.

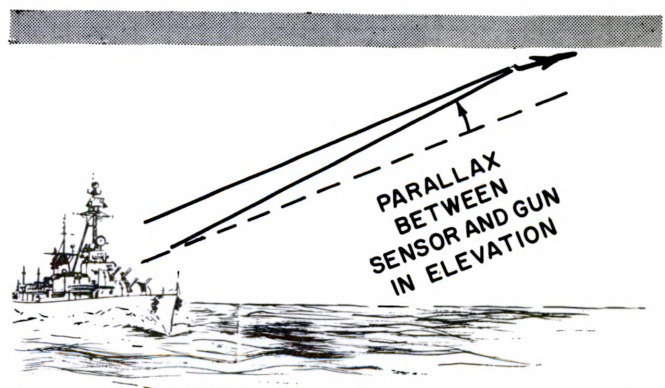
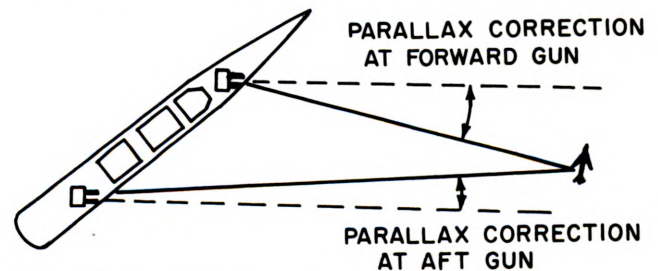
Orientation of the sensor platform is not a serious problem on ground installations, but on moving vehicles it becomes considerably more difficult. For instance, ship-based radar tracking an aircraft will have two factors causing its tracking line to deviate from the line of sight, the target's motion and the ship's motion (pitch, roll, and yaw). Stabilization of the sensor tracking line with respect to ship's motion is usually accomplished through automatic control by a gyro-monitored servo loop.

The fifth factor, the path which the emitted energy will follow to a target and return to its source, is a problem extremely pertinent to sonar systems. Oceanographic charts, combined with stored information on the ocean's properties (salinity layers, etc.), and the effects which prevailing conditions (weather, underwater detonations, etc.) might have on these properties, are all necessary factors in computing the energy path prediction. Energy path prediction is becoming increasingly important to detection systems employing electromagnetic radiation. Atmospheric and meteorological conditions have a pronounced affect upon radio and radar energy propagation. Accurate predictions promote optimum employment of the detection equipment.

Own weapon station location is the equivalent to target location data for a geographically fixed target. In the control of a long range weapon directed at a fixed target, the need for an accurate knowledge of the weapon

system location is apparent. Any error in its location will result in an equal error in determining target position. Own location is provided by charts, motion sensors, and location sensors.

Parallax is a significant problem in target location. A weapon system must correct for differences in train and elevation between different sensors and/or between sensor and weapon. The degree of correction needed depends on the separation in train and elevation between the system components exchanging target location information. Parallax varies directly with the separation of the components concerned and inversely with target range.



Target motion and target maneuver are concerned with target location prediction, which is necessary for contact maintenance (tracking) and weapon intercept geometry calculations. As long as a target maintains constant motion, target motion prediction is simply an extension of successive target locations. However, target maneuver detection is a different matter. Most computers smooth either the input data or the target motion solution to eliminate or minimize undesirable noise from the display or control signals. Unfortunately, the smoothing constants are often sufficient to delay the realization of target maneuver for an appreciable period of time. Some other means of maneuver detection is therefore highly desirable, especially in ballistic weapons where the projectile is not subject to inflight guidance and control. On the other hand, missiles with homing devices are not as easily fooled by target maneuvers. The sensors required for target location are usually of the same type as those used for active target detection. Passive detection systems, while useful in determining target bearing, give little or no information as to target range and are easily defeated by target energy silence. (Passive visual systems are the exception.)

tracking

Target location sensors include those which gather the information necessary to define a target's location, and those which actually lock on and track the target.

Tracking consists of bringing a controlled line (or tracking line) into near coincidence with the line of sight to the target. The target energy reflections or emissions are tracked by optics for visible light energy, by heat sensitive detectors for infrared energy, by automatic tracking radar sets for radar energy, and by sonar sets for acoustic tracking.

Tracking is essential to target interception since it provides the necessary flow of data (either continuous or intermittent) which is so important for target location. It may be accomplished by a sensor which has two operational phases (one, sighting a target, and, two, locking on and tracking it), or by two or more separate sensors, in which a searching sensor detects the target and locates it with sufficient accuracy to supply information to a designated tracking sensor.

Tracking may be periodic or continuous; to accommodate multiple target situations, a system may employ periodic tracking. Several targets can then be tracked at the same time, data from each target being observed and processed during a portion of a time-sharing cycle. A target path is then defined approximately by a series of points on the weapon control system display. Continuous tracking gives continuous plot of the target path and is usually the more exact method. However, continuous tracking is usually required only for high speed and/or highly maneuverable targets.

CLOSED-LOOP SERVO SYSTEMS

Tracking systems normally operate as closed-loop servo systems, referred to as tracking loops. A sensor tracking loop consists basically of an energy sensing device following or tracking a target through continual adjustment of its positioning device. The adjustment is based on constant comparisons of target information feedback (mainly line-of-sight data) with the direction of the sensor axis (tracking line).

The operation of a sensor tracking loop may be manual, fully automated, or only partially automated. In a manually-operated sensor tracking loop, such as one using a rifle sight, a man tracks the target optically, continually shifting his sighting device in accordance with his own judgment. In an automatic system the tracking operation is independent of man. The sensing device is trained and positioned by a servo interpretation of the error signal. In a semiautomatic system man performs some part of the tracking operation, usually interpreting the sensed information. The degree that man participates in a tracking operation depends on the speed, sensitivity, and accuracy of the operation. Speed can be gained at the expense of sensitivity and/or accuracy and vice-versa. Therefore, a series of compromises is necessary to achieve a balance between conflicting operational requirements. Accuracy requirements, for example, are greater for a long range weapon system than for a short range weapon system of com-

parable capabilities. Computing speed requirements, on the other hand, are greater for a short range target than for a long range target since the former presents a more immediate threat.

target association

Targets may be located and tracked in succession by several different tracking systems. These tracking stations may be located at a single weapon station, such as a vehicle mounting both search and fire control radars, or they may be located on separate weapon stations. In the former case, transfer of target information from the search radar to the fire control radars can be accomplished quite easily. When target locations systems are located in weapon stations many miles apart, the problem of target association becomes increasingly difficult. Target association is the problem that multiple tracking stations have in correlating information about specific targets.

All targets must be detected, tracked, and the resultant information passed on to control handling systems. The individual identity of each target must be established by successive tracking stations so that all targets are accounted for. A target must never be allowed to slip through because a station thought it was something else. The importance of target association is readily apparent when one considers the possible ability of one target, moving undetected through a radar screen, to destroy the entire defense potential of the defender.

countermeasures

A target can defeat a target location system in a variety of ways. Any countermeasure which gives false target information, such as chaff, decoys, etc., can be effective since they tend to obscure exact target position and may overload the system with false targets. In these instances, the most effective counter-countermeasure is a highly trained operator or an accurate discerning mechanism capable of distinguishing false targets from real ones.

Jamming techniques are similarly effective because exact target position is lost in the jamming signal. Delay circuits, homing devices, and randomness of output are all effective methods of countering jamming techniques. Vehicles capable of carrying even more complex equipment than that required for jamming may actually transmit false positional information that may be picked up and interpreted as energy reflections by the system attempting to locate it. This method is particularly effective against detection systems relying on the Doppler frequency shift. Since this countermeasure involves not only recognizing a signal, but also computing and transmitting another signal, it cannot as yet be accomplished instantaneously. Therefore, coded or random pulse trains are effective counter-countermeasures at present.

Spoofing or electronically simulating targets is also employed to overload and fool target location systems.

Again, coding or randomness of outputs can successfully defeat this form of countermeasure.

It is essential that a system be flexible enough to handle all of the countermeasures it can expect. Utilization of alternate sensing devices, either of the same or of a different type, can insure this flexibility. It is also imperative that the system recognize false targets in their initial stages. Once it begins tracking them, their purpose has been accomplished and if there are enough of them the system will overload and break down. System overloading may also result from tracking an abundance of real targets. The high-traffic tracking problem can be solved through multiple sensors and/or periodic tracking. Splits and merges are conditions inherent to the multiple target problem and add to tracking difficulties. Splits occur when a contact detected by the sensor as a single target splits into two more discernable targets. Merges occur when several targets come together so that they appear as one target to the detection system. Considering these problems and the possible countermeasures a complex target might employ, a system could easily break down if it did not make use of channels and queues. In a system with multiple sensors, the channels are the individual tracking systems and their related data processing (computing) systems. Since a tracking device can usually follow only one target at a time, it will also employ channels in its data processing when it is assigned more than one target. Inputs are handled usually in order of their potential danger or in order of arrival. When the number of inputs is greater than the number of channels, a queue will form which is a function of the target arrival rate. However, system breakdown is still possible since the system approaches overload and breakdown as this waiting line gets longer. In this sense, breakdown does not mean that the system actually ceases to func-

tion, but, rather, that it is no longer furnishing an appreciable portion of the target information needed to handle all of the targets. The ability to pick out and ignore false targets is therefore of prime importance in handling multiple target situations.

man-machine tasks

Allocation of target location tasks between man and machine is an important factor in system operation. Those tasks usually reserved for man are:

- 1) Selection and designation of the target data input source
- 2) Selection and control of machine operating modes
- 3) Determination and execution of own vehicle maneuvers to optimize problem geometry
- 4) Manual input of data not otherwise available to the machine (i.e., a target maneuver detected at the sensor might be used as a basis for sensitivity control to quicken machine response)
- 5) Machine control and display monitoring.

Machines are usually allocated the following tasks:

- 1) Storage of computational programs
- 2) Three dimensional target location and motion computation, using any or all available data
- 3) Stabilization of measured data
- 4) Computation of energy path and associated data corrections
- 5) Continuous target position generator and prediction
- 6) Parallax computations
- 7) Own ship position generation
- 8) Signal analysis for target maneuver detection
- 9) Display of target and own ship position
- 10) Computation and display of target data analysis.

WEAPON SELECTION PHASE

The weapon selection phase is defined as the selection of the best available weapon in a launching vehicle's armament to inflict optimum target damage. Although the term target damage implies actual physical destruction, this is not always the case because the task of a weapon system may be psychological rather than

destructive. The weapon system is designed or organized with proposed objectives in mind. Once the objective is determined, the system is committed. Therefore, it must be chosen very carefully to cover all practical conditions which could deter mission accomplishment.

information requirements

The information factors necessary to accomplish weapon selection are:

- | | |
|-------------------------------|------------------------|
| 1) Target classification | 5) Mission |
| 2) Target location and motion | 6) Weapon Capabilities |
| 3) Geography | 7) Weapon availability |
| 4) Own force safety | 8) Kill Probability |

Target classification and location determine weapon requirements and destructive potential and range velocity, etc. of the target. These information factors are available from preceding operational phases.

Geography is important in determining any environmental limitations on the proposed weapon choice. In hilly or mountainous terrain, the effectiveness of flat trajectory weapons may be seriously reduced. Own force safety is the consideration of the proposed weapon's flight path and destruction capabilities relative to known deployments of friendly forces. Guided missiles, for instance, must not be fired in directions where their expended boosters would fall on friendly forces. Knowledge of the mission contributes to weapon selection because it determines which attack is most urgent and the required degree of stealth or surprise, which is obtainable only with specific types of weapons.

Weapon capabilities are determined generally by the design of the weapons systems and their installation in the tactical vehicle. Tactical employment of these weapons is a function of the battle and the vehicle's mission. Information as to weapon availability can be obtained by simple accounting methods and must be supplied continuously to the person responsible for weapon selection.

Kill probability is the evaluation of the usefulness of a weapon in the prevailing circumstances. The relative kill probability of the various weapons available should be considered before selection is made. Of all the factors necessary for weapon selection, only kill probability entails any special requirements. The information requirements of the other factors are satisfied either by routine preoperation briefing or instructions during the attack.

Kill probability, however, is not easily obtainable. The kill probability of a given weapon is a function of:

- 1) Lethal radius and/or the acquisition range of the weapon
- 2) Calculated miss distance of the weapon (from the aim point)
- 3) Target evasion capability.

Lethal radius and acquisition range can be determined from weapon design and test data. Miss distance can be formulated as a function of solution quality (based on computed range, computing time, solution, stability, quantity and nature of input data), known error distribution in sensed data, computing accuracy, and known weapons dispersion. Target evasive capability can be computed from estimated target maneuver capability and the time delay from data measurement to weapon arrival at aim point. The calculations are sufficiently complex to make machine processing the only practical approach.

weapon characteristics

Initial weapon system design is essentially weapon selection for a hypothetical target complex. This involves evaluating the vulnerability of typical targets and realizing that targets possess varying degrees of vulnerability to particular weapons, depending on their physical characteristics and environment. To achieve optimum results, the system employing a particular weapon must use it against the same general class of target which dictated its design.

In a particular tactical situation, weapon selection is limited to the weapons on hand. It is necessary to select the weapon which is best able to perform the particular function required. This involves consideration of the weapon destructive capabilities, target velocity, and target range. The weapon selection phase is dependent therefore on the target classification and target location phases. The weapon is selected which has the highest probability of destroying the target.

weapon kill probability

The possibility of a weapon succeeding against a particular set of target circumstances is referred to as its kill probability. The kill probability of various weapons available relative to the situation at hand must be considered before selection is made.

When the weapons at hand do not have sufficient kill probability against a given target, the weapon system may be able to improve the situation by maneuvering the weapon station. For example, maneuvers which operate to reduce the range or to reduce missile-target relative motion can effectively improve kill probability. If the kill probability of a single weapon, or a single missile from a particular weapon, cannot be made sufficiently high, it may be desirable to fire several missiles at the target simultaneously to obtain a high overall kill probability. The firing of several missiles at once or in rapid succession is called salvo fire. If a salvo of n missiles, each with a kill probability, P_k , is fired at a target, and the cumulative damage inflicted by successive warheads is ignored, the salvo kill probability is approximately:

$$P_{\text{salvo}} = 1 - (1 - P_k)^n$$

Thus, if a single missile has a kill probability of .5, a two-round salvo will have a cumulative kill probability of .75 and a three-round salvo a cumulative kill probability of .875. The choice of the number of missiles in a salvo depends upon the desired cumulative kill probability, weapon availability, and launching facilities.

In cases where weapons of only limited kill probability are available, the weapon system must often settle for less than total destruction of the target. If it must engage the enemy, its weapons can be employed to harass or threaten, even though little actual damage is inflicted. Designation of the weapons to handle specific targets is usually under the control of personnel trained in evaluating target location and classification data with regard to the weapon capabilities of their own vehicle. It is their responsibility to know what weapons are available and whether they are functioning properly. The speed with which proper weapon designation can be made is a major factor in overall weapon system effectiveness. The communication links between the weapons designer and the elements of the system furnishing him with the information he requires play an important part. The faster the target data is processed and transmitted, the quicker he can make his decision. He must also be in constant communication with the various

weapons which constitute his system, not only to obtain instantaneous response to his decisions, but also to be constantly informed as to the availability of each of the system's weapons.

man-machine tasks

In the weapon selection phase, machine capabilities are required for the continuous calculation of kill probability. In addition to computing the kill probability for the proposed weapon, machines are also required to furnish a display of the tactical situation and of the kill probability computation results.

The weapon selection tasks reserved for man are:

- 1) Assimilation of operational and doctrinal instructions
- 2) Data storage, by training, of weapon capabilities and characteristics
- 3) Survey of the tactical and geographic situation
- 4) Machine control and display monitoring
- 5) Weapon selection decision

WEAPON LAUNCHING PHASE

The weapon launching phase concerns the launching of a missile into the desired flight path in a safe and efficient manner. Because the launching phase is the final phase in weapon system employment, it is concerned not only with proper performance of the launching system, but also with missile flight and terminal ballistics, attack evaluation, and preparation for another attack. Safety is of prime importance during the launching phase because the handling and launching of missiles is always dangerous.

information requirements

The information factors which are pertinent to this phase are:

- 1) Weapon readiness and speed of employment
- 2) Weapon characteristics and capabilities of shifting from target to target
- 3) Target environmental situations
- 4) System kill probability
- 5) Own force safety
- 6) Postfiring launcher condition and status

As this information is processed, the material system responds to the decisions to transfer, load, and launch missiles. Adequate information is essential; however, care must be exercised to insure that the information channels do not reach saturation. This saturated condition would only result in slowing the response of the material system.

During the launching phase, it is necessary to make the selected weapon and its associated missiles ready for launch at a designated target in a minimum amount of time. It is towards this rapid initial employment of a selected weapon that much of the design effort in the weapons field is directed. It is obvious that the weapon must be brought into action against a target before the enemy can accomplish his mission, or the weapon will be of little use.

Of equal importance is the ability to shift from one target to another, thus allowing one launcher to engage multiple targets. Speed of employment must be within the requirements imposed by the characteristics of the target. The multiple target situation and low kill probability of some missiles necessitates some launching systems to launch many missiles at a high rate to achieve the desired kill. This continuous operation of a launching system may dictate the need for two or more weapon control systems to guide missiles against multiple targets. In some cases, if the number of weapon control systems is limited or the launching system is slow, it may be necessary to launch missiles from a number of launchers to increase the kill probability by salvo fire and thus reduce the time a weapon control system is committed to a particular target.

selection of time to fire

It is necessary to evaluate the system kill probability in deciding what, how many, and when to fire. Generally, the missile should be launched at such a time to allow intercept of the target at the maximum effective range. This will provide time for another attack if the first missile should fail, and will keep the enemy from closing and delivering his ordnance. In some situations, where the missile kill probability is low or the missile supply is limited, the range to the target may be allowed to decrease so that the probability of a kill will be increased. Many large weapons systems have weapons with widely varying range capabilities. As a target penetrates into the effective zone of each weapon, it is brought into action, thus increasing the accumulative kill probability. Thus, inherent to the launching phase is the need for rapid, accurate evaluation of the success of an attack in order that a second attack can be initiated immediately if the target is still a threat. It should be noted that complete target destruction is not necessary. If the target is damaged in such a way so that it ceases to be a threat, the target has been successfully engaged. The commander, in deciding to attack a second time, must always consider his mission, the missile supplies available, other targets which may provide a greater threat, and the chance of success if he should attack again. To assist the commander in making this decisions, he must not only be kept informed of the terminal effect, but must be up to date in the status of the launchers themselves. Before launch the information provided should state when a missile can be launched, and what type missile and warhead will be available.

Immediately after launch, information must indicate whether the launch was accomplished successfully. The information must indicate if there is a need to fire immediately a second missile, or if a dual, hang fire, or malfunction demand an immediate second attack.

launch functions

To fulfill the operational requirements imposed during the launching phase, a launching system performs certain specific functions:

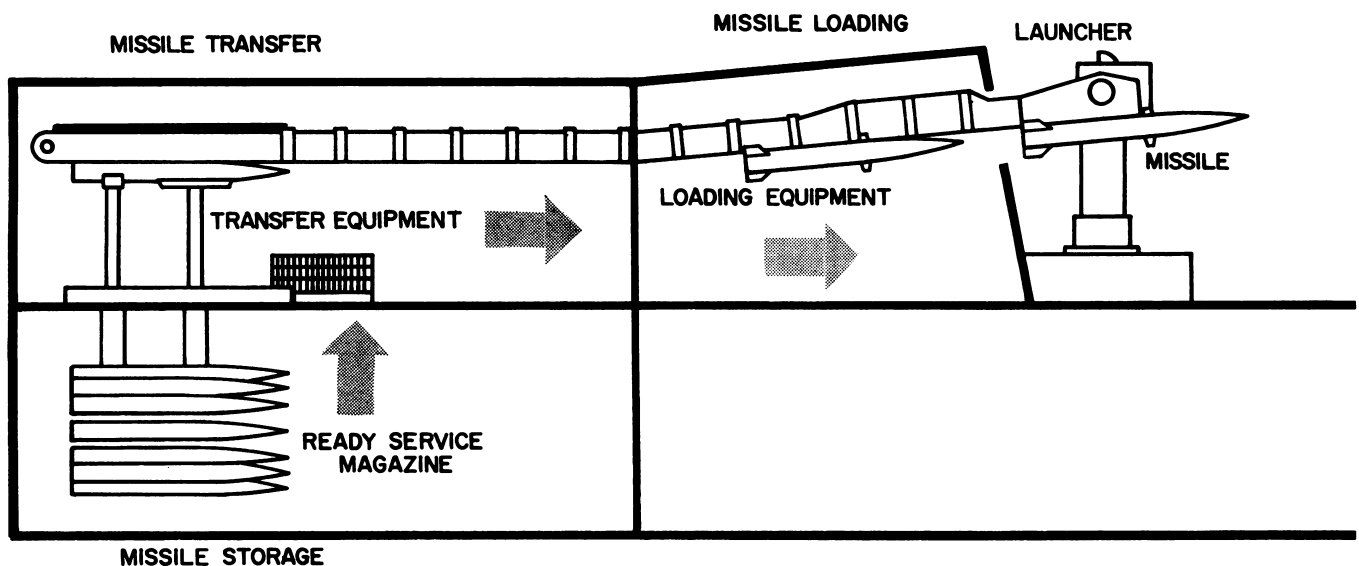
Missile Storage
Transfer
Loading
Launching

Associated with these functions are functional components of the launching system.

The magazines and ready service stowage provide for the safe, protected storage of a supply of missiles needed to engage the enemy. This supply is dictated by space limitations and types of targets expected.

Transfer is accomplished by moving the missile from storage areas to the launcher. The speed of transfer is dictated by material limitations and the demands of the weapon system design. If a single transfer channel does not satisfy the rate of fire requirements, two or more transfer channels may be employed.

Loading requires that the missile be moved from the transfer channel to a position on the launcher. Techniques for increasing rates of fire may include multiple loaders supplying a single launcher or even a single loading device loading numerous launchers. During loading and transfer, flight preparation is generally accomplished.



positioning of missile

The fundamental unit of a launching system is the launcher which accomplishes flight initiation of the missile. Uncontrolled missiles receive all of their guidance during the few moments they are on the launcher. Considerable precision in launcher orientation is desirable with this type of missile to achieve effective employment without undue expenditures of missiles.

Launchers which are designed to handle controlled missiles generally have a greater latitude in positioning the missile. Precision of orientation at launch does not dictate the resulting kill probability since the missile travels along its variable or controlled path to the target. This is not to say that the launcher position on firing can be ignored. Most missile systems have correction and attitude control limits within which they must operate. The launcher must put the missile within these boundaries at launch or the missile's guidance systems (either internal or external) may not be able to compensate for the discrepancy and gain control of the missile.

man-machine tasks

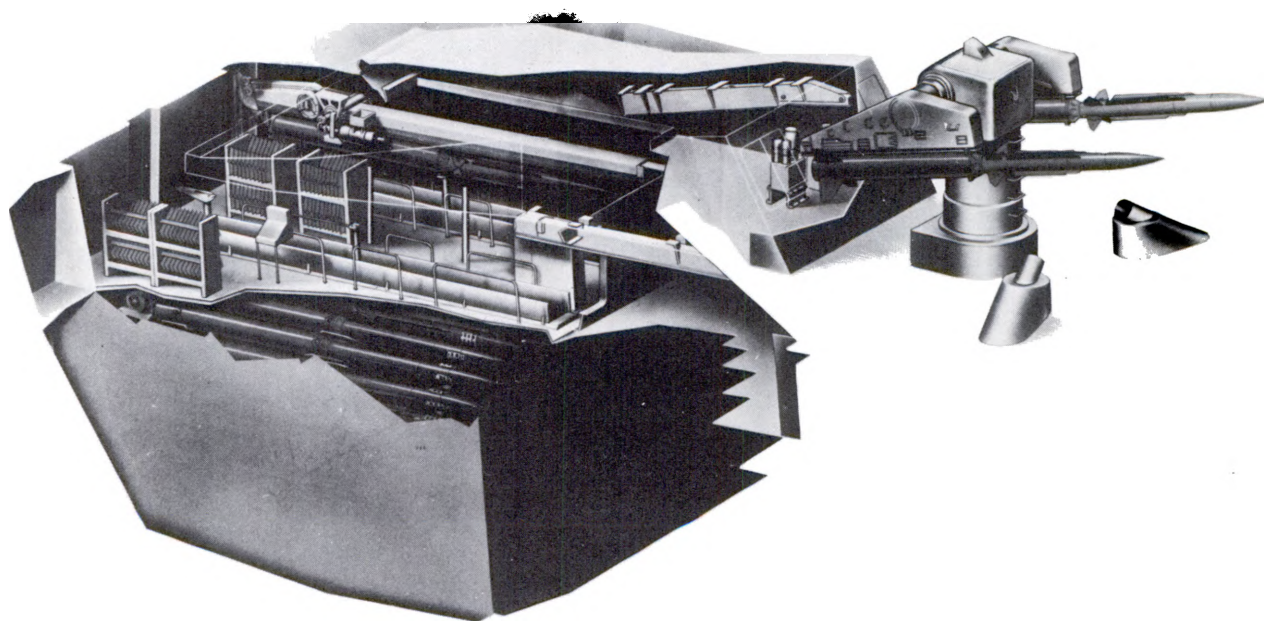
The following is the allocation of weapon launching tasks for man and machine.

Those tasks commonly reserved for man are:

- 1) Monitoring of safety factors
- 2) Selection of firing time and initiation of weapon firing cycle
- 3) Conduct of weapon launching procedure, including override control of automatic sequencing or machine powered operations when needed for reasons of safety or problem deterioration, or for a change in the tactical situation
- 4) Evaluation of attack success and decision as to future action

The tasks generally handled by machine are:

- 1) Display of kill probability and selected weapon limits
- 2) Power assistance for all difficult, critically timed, or remote physical operations
- 3) Sensing and display of inaccessible and remote physical conditions
- 4) Automatic sequencing of critical or complex switching and monitoring-switching cycles.



WEAPON DIRECTION PHASE

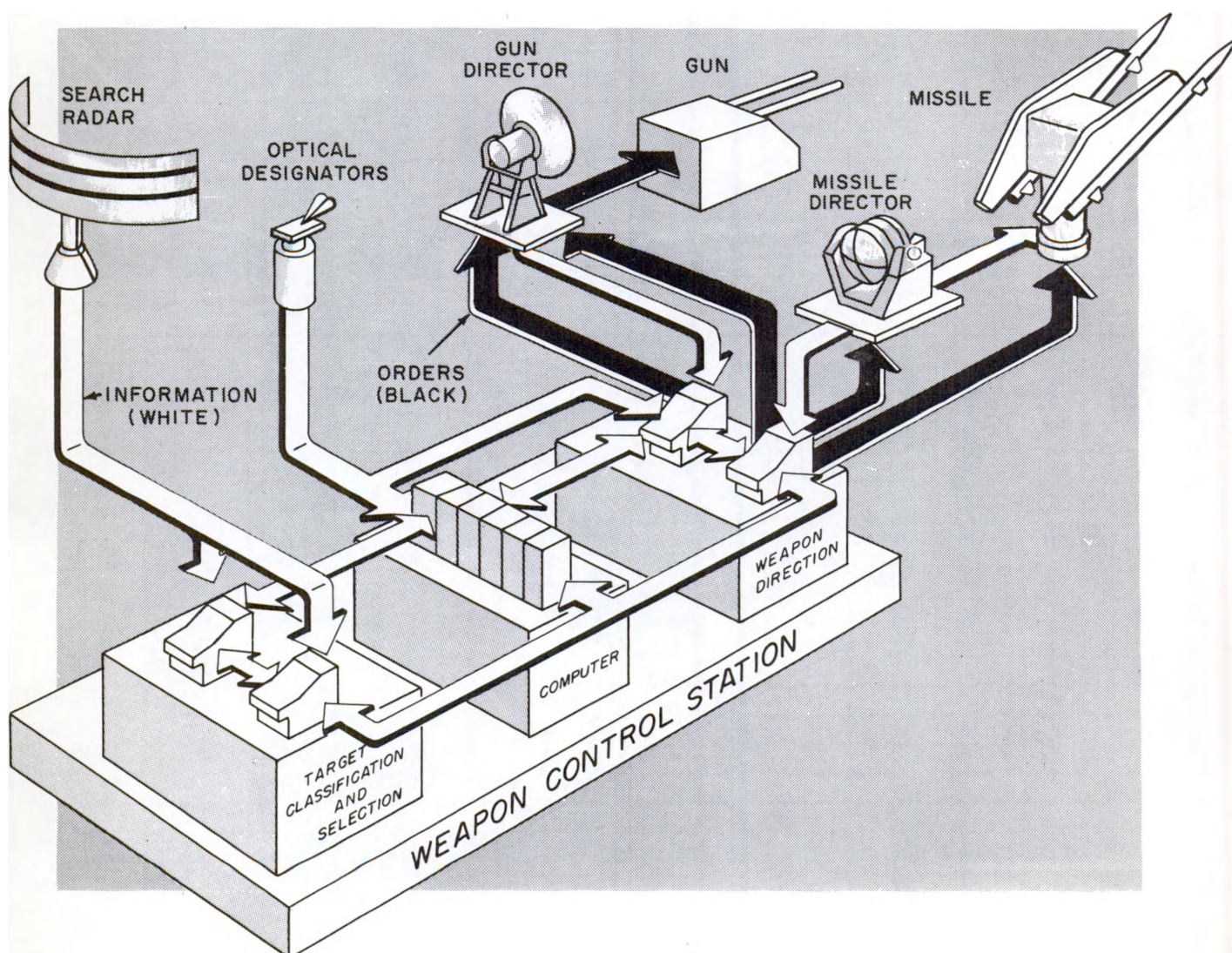
The weapon direction phase is defined as the acquisition of the target by the weapon control system, and the gathering and processing of target and related information by the weapon control system to generate weapon control orders and information displays. Having data on the target, it is necessary to determine the direction in which the missile should be aimed to ensure

interception with the target. From the data on target position and velocity and weapon system environment, a prediction angle is computed which is the angle that the weapon line must lead the line of sight in order to intercept the target. The missile can be launched whenever the weapon line is correctly positioned.

information requirements

The information necessary for the successful completion of this task involves knowledge of:

- 1) Weapon ballistics - kinematic lead factor and jump effects
- 2) Weapon control characteristics
- 3) Target location and motion
- 4) Weapon alinement
- 5) Weapons platform reference
- 6) Weapon platform initial conditions
- 7) Weapon terminal guidance capabilities and lethal radius
- 8) Weapon path
- 9) Target location and motion during weapon travel



The processing of these information factors is performed by information processing devices, such as analog and digital computers. These machines far surpass man in speed and accuracy when the parameters with which they are to operate are distinctly defined. The first two information factors, weapon ballistics and weapon control characteristics, are the starting point in determining ballistic equations, the solution of which results in weapon direction orders which establish the weapon flight path. Target location and motion information constitute the input data for the ballistic equations. The ballistic equations transform the input data into weapon orientation and range orders in the form determined by the weapon characteristics. The equations normally account for all weapon path perturbations which can be measured or calculated. Kinematic lead involves a determination of target motion and lead angle in the positioning of the launcher along the weapon line and not the line of sight. Jump is an effect of vehicle movement during launch; trajectory deviations as a result of this effect must be accounted for. Weapon alinement involves the orientation of the weapon and weapon platform, relative to the fire control reference of the delivery vehicle. The weapon control problem will generally involve three elements, the target, the delivery vehicle, and the weapon itself, one or all of which may be in motion. The motions pertinent to solving the weapon control problem are those which the elements involved have in relation to each other. To determine these relative motions, therefore, it is essential that all three be located within the same reference frame.

The reference frame chosen may be fixed in space (inertial), in which case the motion of Earth must be considered, or it may be fixed with respect to the Earth or the air mass. The reference frame may also be fixed in the delivery vehicle. This is usually accomplished by establishing fixed planes that move with the vehicle but are effected by the vehicle's rotational motions. Vehicle-based reference frames are commonly employed on ships and aircraft.

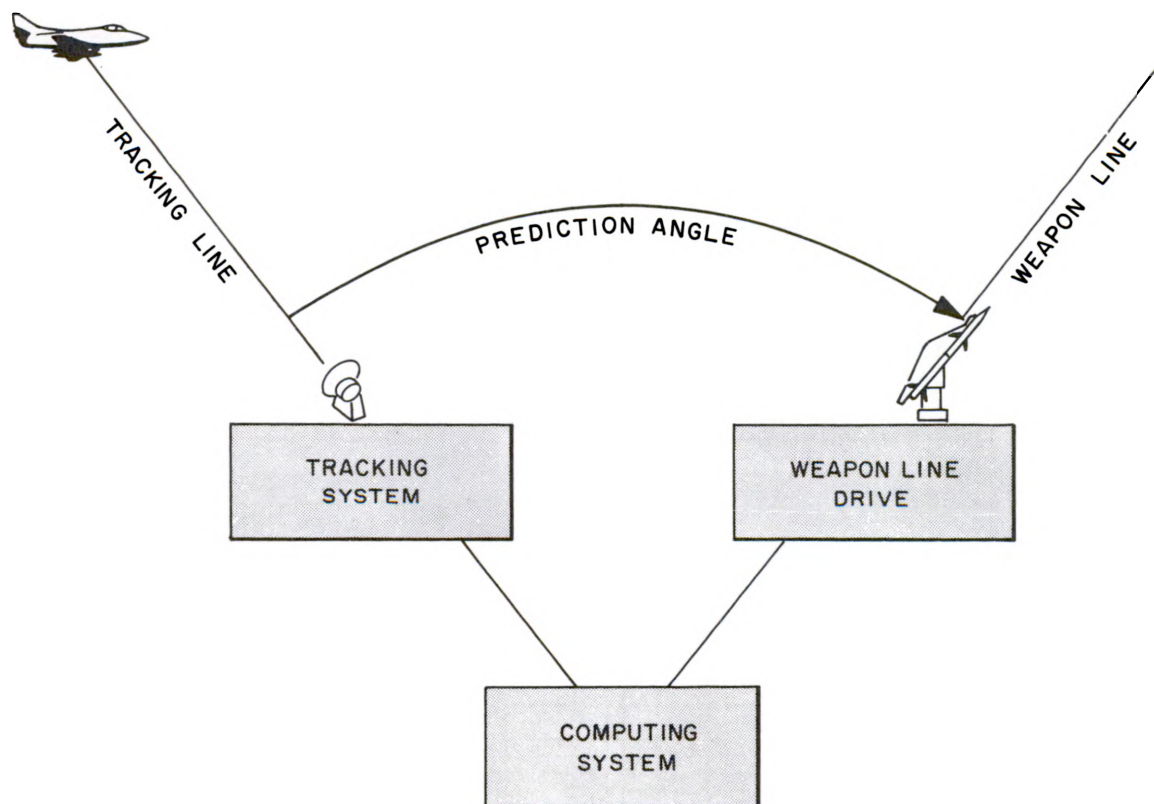
Weapon terminal guidance capabilities and lethal radius are the data which must be considered in controlling the weapon's performance at the end of its ballistic path.

accuracy required

The accuracy required of the weapon direction phase varies inversely with the lethal radius of the weapon involved and directly with target range. A weapon having a large damage volume, such as one employing a nuclear warhead, does not have to be detonated as close to a target as one employing a conventional warhead to accomplish the same damage. Range, on the other hand, effects accuracy directly. Errors insignificant at close range are critical at long range.

speed

The speed of weapon direction depends on the type of target or targets for which the system is designed. Generally, rapid response of the system is of the utmost importance, and systems are designed to have the greatest speed commensurate with the accuracy required to achieve the desired terminal effects.



types of operation

Weapon direction can be manual, aided, or automatic. In a manual system the operator aims his weapon according to his own judgment and uses his own power to achieve target interception. In most systems, however, because of the weapon size and the need for precise positioning, he is aided by electromechanical means to position the weapon correctly. The data which is required for successful target interception generally entails computations that have speed and accuracy requirements far beyond his capabilities. Computers, therefore, are indispensable to most weapon systems. A completely automatic weapon direction system is one in which all the necessary operations are normally performed independently of human control. By continuous processing of the input data in the computing system, the weapon control system constantly directs or repositions the weapon line. Thus, missiles can be launched from the selected weapon at any time the control officer chooses. When the designated weapon launches uncontrolled missiles, the launcher must be continuously and accurately positioned so that a sufficient number of missiles can be launched to achieve the desired target damage. If the missiles from the selected weapon can be controlled after launch, the weapon direction phase still continues until target interception. In the latter case, however, postlaunch weapon direction is concerned with the missile and not the launcher.

uncontrolled missiles

An uncontrolled missile launched into free flight is committed at the instant of launching to a particular flight path because no further control can be exercised by the weapon control system. Therefore, accuracy requirements are quite high and all weapon direction junctions and decisions pertaining to any specific missile are made at the weapon station prior to launching. The path followed by any propelled object subject only to its initial velocity, its initial attitude, and the forces of nature present (gravity, wind, and air resistance) is considered to be ballistic. Under this classification falls the paths followed by missiles propelled by impulse and depending for guidance solely on conditions imparted to them just prior to release. Arrows, bullets, artillery projectiles, and bombs are all examples of missiles that follow a purely ballistic path after release. Missiles which are propelled by reaction propulsion systems and which are without control guidance during flight, such as free rockets, follow ballistic paths after engine cutoff.

controlled missiles

A missile whose flight path is controlled after launching is considered to be guided. Internal equipment may sense its deviation from the prescribed path and operate to correct it, or it may be commanded externally to make certain changes in its path. Many systems such as those in a long range ballistic missile employ a com-

bination of both guided and unguided trajectory periods. Guidance and reaction propulsion are usually employed in the initial and/or terminal stages of flight, while during the intermediate portion the missile will follow a ballistic path.

Every missile guidance system consists of an attitude control system and a path control system. The attitude control system exercises control of the missile in roll, pitch, and yaw. It functions to maintain the missile in the desired attitude on the ordered flight path. The attitude control system of a missile operates essentially as an autopilot, damping out flight perturbations that tend to deflect the missile from its ordered flight path. The function of the path control system is to determine the flight path necessary for target interception and to generate the necessary orders for the attitude control system to follow.

Thus, the missile guidance system is essentially a weapon control system and is inherently associated with the weapon direction phase. Therefore, when a weapon with controlled missiles is employed, the weapon direction decisions continue to be made after the missile has been launched. The weapon direction (weapon line orientation) decisions may be made at the weapon station or in the missile itself, depending on the nature of the guidance system.

guidance systems

Guidance systems are usually called by the name of the path control system since many missiles use the same type of attitude control. There are essentially eight basic guidance systems:

- 1) Preset
- 2) Inertial
- 3) Terrestrial navigation
- 4) Celestial navigation
- 5) Radio navigation
- 6) Command
- 7) Beam rider
- 8) Homing

The first four systems are self-contained guidance systems in which the missile in each case is made to fly a path calculated before launch and is kept on this predetermined path by equipment contained wholly within the missile itself. In this type of missile all weapon direction decisions are made at the weapon station prior to launch. The missile guidance system merely gathers information and generates flight control orders to keep the missile on the previously selected flight path. No decisions are made by the guidance system as to the selection of new flight paths to the target, nor are any new flight paths generated. Systems of this type are especially useful in guiding surface-to-surface missiles. A principal advantage is that there are no existing countermeasures. A principal disadvantage is that accuracy is generally poor because of instrument errors. As a result, these systems, unless combined with another system, are generally limited to area targets.

PRESET

In a preset guidance system a predetermined path is set into the missile prior to launching and cannot be adjusted during missile flight. The principal device used in the preset missile is a programmer. Predetermined operations are programmed into this equipment causing the components in the missile to perform certain definitive functions during missile flight.

INERTIAL

An inertial guidance system is one designed for a predetermined path. The path of the missile is adjusted after launching by the guidance system. Inertial sensors are located within the missile which are independent of any outside information. The main components are accelerometers to detect and measure acceleration in three directions. The accelerometer outputs are integrated twice to obtain the lateral and vertical distances that the missile is off from its predetermined flight path. By doubly integrating the missile acceleration along the desired flight path, the distance the missile travels is found. The path control system compares the actual missile position with the desired position along the intended flight path, and generates orders for the attitude control system to keep it on the required flight path to the target.

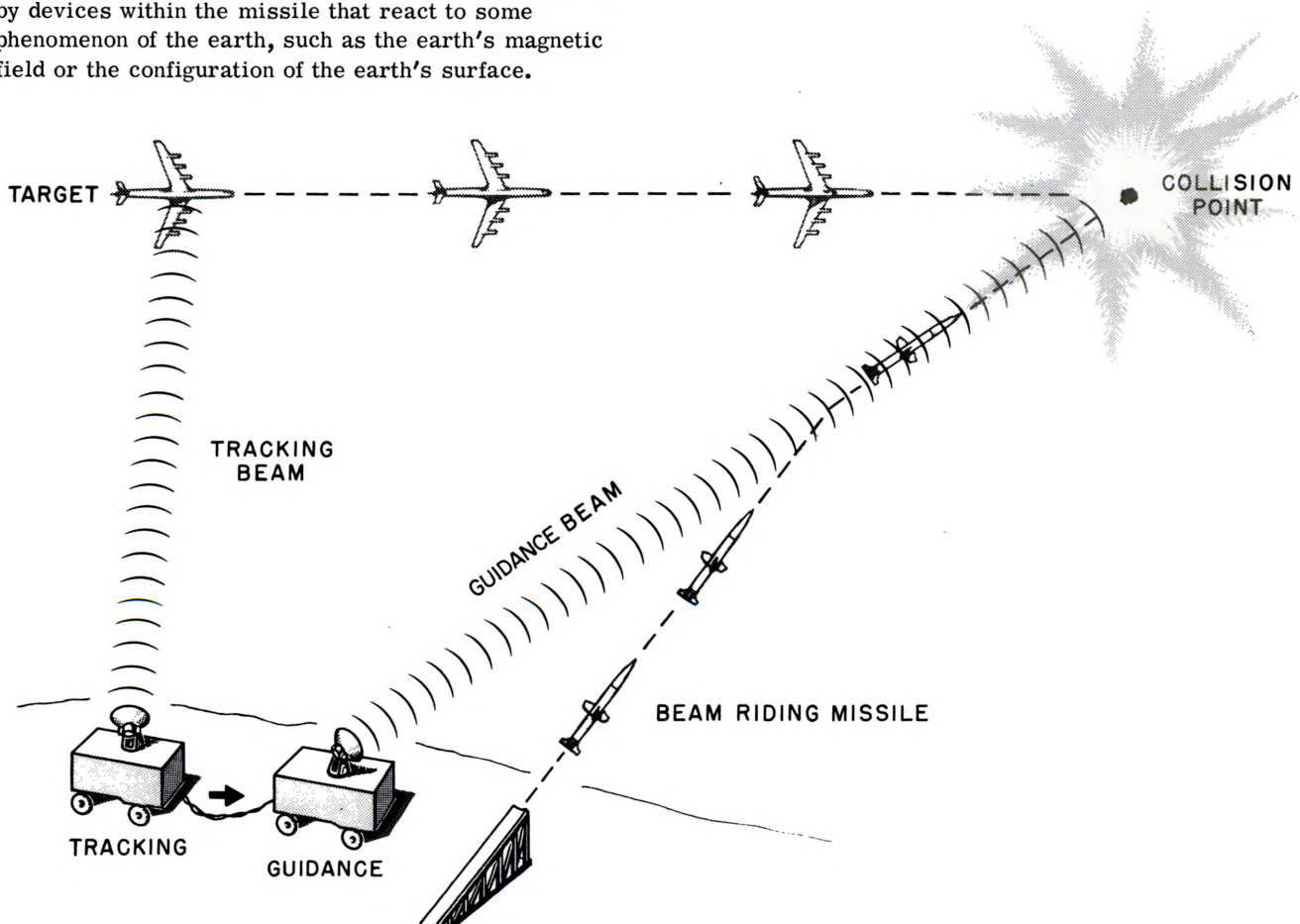
TERRESTRIAL

A terrestrial reference guidance system is one in which the predetermined path can be adjusted after launching by devices within the missile that react to some phenomenon of the earth, such as the earth's magnetic field or the configuration of the earth's surface.

CELESTIAL

A celestial navigation guidance system is a system designed for a predetermined path in which the missile is adjusted by use of continual celestial navigation. The system is based on the known apparent positions of celestial bodies with respect to a point on the surface of the earth at a given time. Such a system is highly desirable for long-range missiles since its accuracy is not dependent on range. The missile must be provided with a horizontal or vertical reference to the earth, automatic star tracking telescopes to determine star-elevation angles with respect to the reference, a time base, and navigational star tables mechanically or electrically recorded. A computer continuously compares star observations with the time base and navigational tables to determine the missile's present position, and the proper signals are computed to steer the missile correctly toward the target.

The last four types of guidance systems listed, radio navigation, command, beam rider, and homing, all depend on signals from a source external to the missile to accomplish target interception. As a result, these systems are susceptible to transmission noise and enemy countermeasures. These systems may adapt themselves to a high degree of accuracy. Generally, they are capable of correlating the position of the target and the missile to correct the flight path and insure target intercept. Through the use of counter-countermeasures, satisfactory kill probabilities may be achieved.



TASKS AND PHASES

RADIO NAVIGATION

A radio navigation guidance system is a system which maintains a predetermined path by adjustments in the missile governed by external radio signals. The simplest method is one in which the missile, by means of a directional antenna, maintains a predetermined bearing with respect to two radio transmitters. Altitude can be obtained from an altimeter, and range may be determined by adding a third transmitting station. Loran and Shoran are examples of radio navigation systems of a much higher order of complexity.

COMMAND

A command guidance system is one in which the missile path is not preset but is determined during flight. A control station obtains missile and target data and directs the missile attack by some data link, such as radio radar. In the system illustrated the target is tracked by one radar and the missile by the other. The computer continuously compares these relative positions and determines new missile flight paths. Radio commands which cause the missile to change its path are sent to the missile by the computer through the missile tracking radar beam.

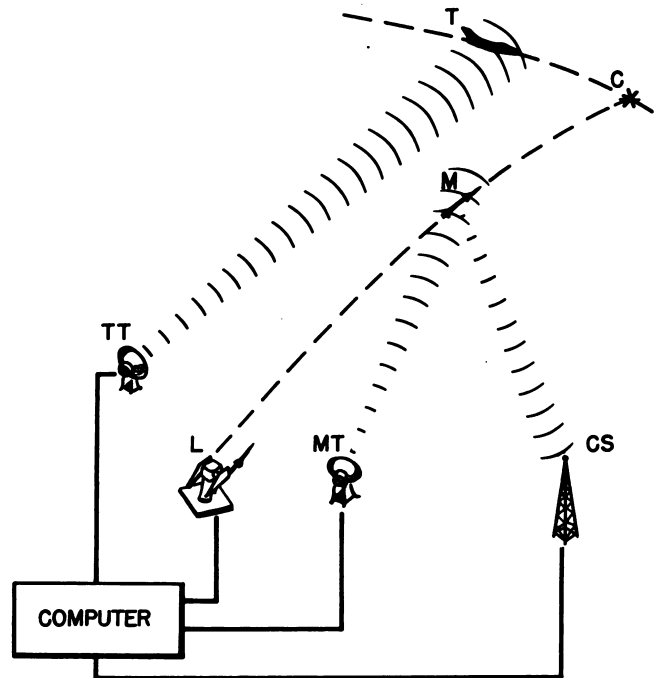
BEAM RIDER

A beam rider guidance system is one in which the missile seeks out the center of a directed energy beam. Radar is the most useful type of beam. Beam rider systems employ either one radar which the missile follows as it tracks the target, or two separate radars, one of which tracks the target while the other guides the missile. In the latter systems the two radars are coordinated by means of a computer. The accuracy of beam rider systems decreases with range because of the spreading of the beam. Thus, they are valuable primarily against short range aircraft targets. A number of missiles can be launched into the beam at once, and if the targets are not too widely separated, the beam can be shifted to a new target with missiles already in the beam.

HOMING

A homing system consists of a seeker or scanner in the missile which automatically keeps locked on or pointed at some special characteristic of the target. The target characteristics of interest are: light, radio, radar, infrared, sound, and magnetic field. Homing systems are generally classified as active, semiactive, or passive. In active homing systems, the missile carries the equipment required to illuminate the target, and the system is independent of an external agent. In a semiactive homing system, some agency outside the missile, such as a surface unit or aircraft, illuminates the target and the missile senses the illumination. In passive homing, the missile homes in on energy radiated by the target, such as heat from a jet exhaust.

Many of the foregoing guidance systems may be combined to utilize the advantages of each system. For example, a surface-to-air missile may use preset guidance during the initial or launching phase to orient the missile in a radar beam, rider guidance during the midcourse phase, and homing guidance during the terminal phase; when the beam rider accuracy decreases. A surface-to-surface long range missile might utilize preset guidance during the launching phase to orient the missile with the target, celestial navigation during the midcourse long-range guidance phase, and homing guidance to enable the missile to pinpoint the target in the terminal phase.



MAN-MACHINE TASKS

The allocation of the weapon direction tasks is generally as follows:

Man is usually responsible for:

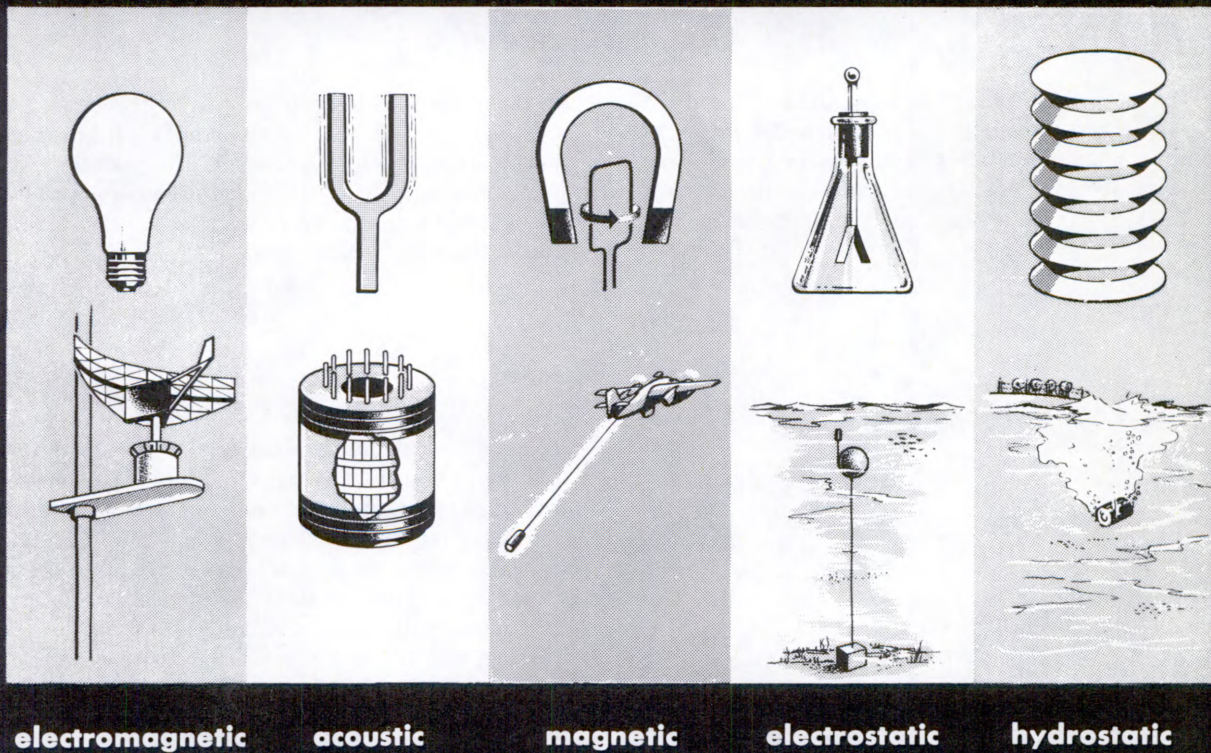
- 1) Determination and manual input of weapon terminal guidance characteristics
- 2) Selection and control of machine operating mode (as a function of the weapon selected)
- 3) Machine control and display monitoring, including override control of weapon orders where warranted by safety considerations or problem deterioration

The tasks normally assigned to machines include:

- 1) Storage of ballistic parameters and ballistic computation programs for each weapon to be handled
- 2) Computation of weapon aim point (weapon line direction)
- 3) Computation of weapon control orders as required by the specific weapons being handled
- 4) Sensing and display of weapon's response to orders
- 5) Computation and display of weapon capabilities with respect to target position and motion

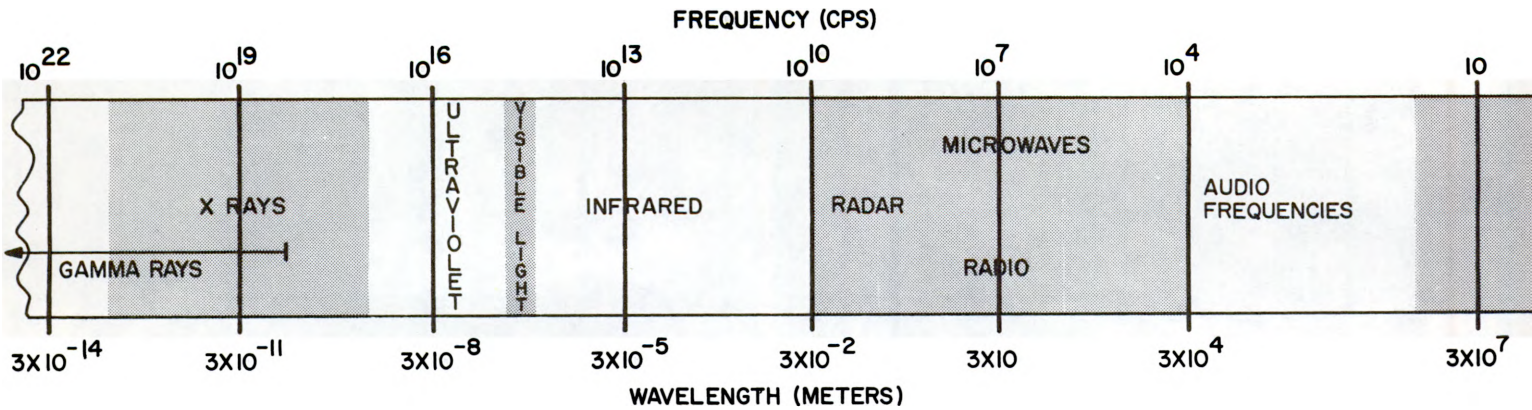
Introduction to

SENSORS and DETECTION SYSTEMS



A primary requisite of a weapon system is that it have the ability to sense or detect the existence of a target. This requires the use of a detection system capable of sensing a characteristic of the target that differentiates or identifies it as such, and furnishing this information to other components of the weapon system. The characteristic of a target that is simplest to sense is the energy form it emits or reflects. This may be either visible light, heat, sound, or electrical energy. Often the electrostatic, hydrostatic, or magnetostatic "signature" of a target is utilized to detect its presence. Target sensing requires that an energy form from the target reach the sensor of the weapon system. This energy is radiated, reflected or perturbed by the target and is sensed by the weapon system sensory or detection devices. Some of the most important energy forms for sensing agents are: 1) the electromagnetic spectrum, 2) acoustical energy, 3) magnetostatic fields, 4) electrostatic fields, 5) hydrostatic pressure.

electromagnetic spectrum



The electromagnetic spectrum is a frequency classification of all electromagnetic energy. In order of increasing frequency, the spectrum includes long range radio, shortwave radio, radar, infrared, visible light, ultraviolet, X-ray, gamma-ray and cosmic ray energy.

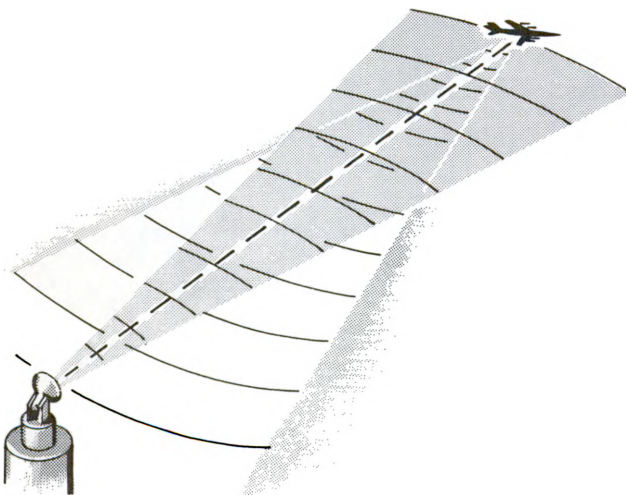
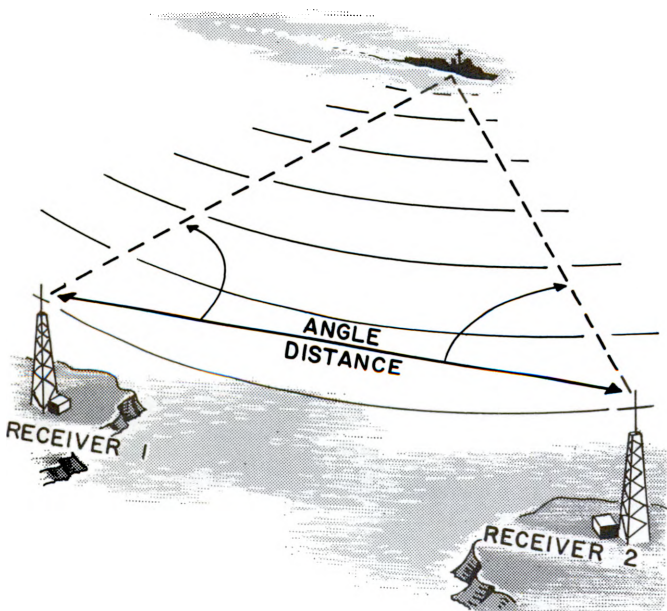
All these energy forms differ only in frequency. However, each portion of the spectrum has distinguishing characteristics and properties. Those portions of the electromagnetic spectrum most often employed for weapons system sensing are radio, radar, television, infrared and visible flight.

radio

Radio waves, at the low frequency end of the spectrum, are used as carriers of information, as navigational aids, or as detectable beams of energy that may be intercepted by a chain of receiving stations operating from fixed or movable locations that can home in on the transmitting signal and locate the position of its source by simple trigonometric means. Radio direction finders (RDF) and radio range detectors are radial systems, that is, the information content of the received energy gives the operator straight lines (radials) of bearing to the transmitter. One radio direction finder receiver can establish direction to a source while two or more receivers can establish direction and range by triangulation.

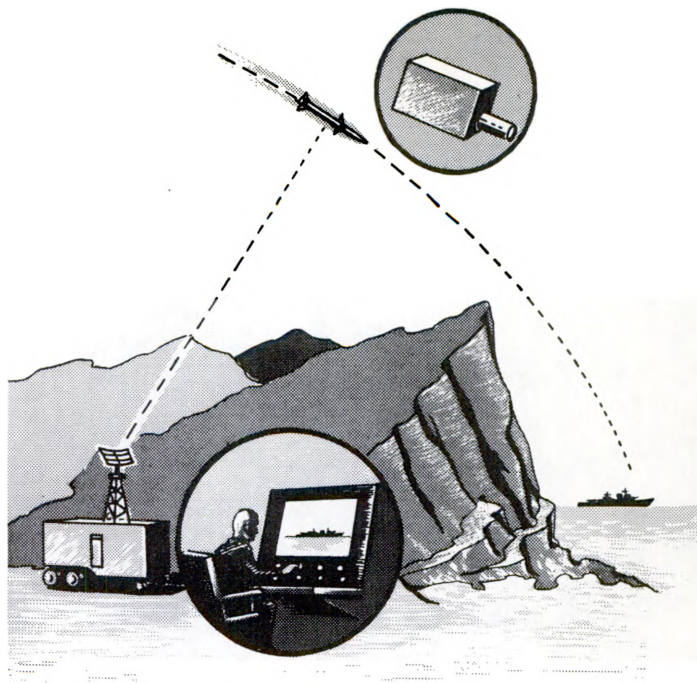
radar

Radar detection depends upon the reflection of radio waves from an object. The word radar, derived from radio detection and ranging, signifies a method of employing the property of a target to reflect portions of the transmitted electromagnetic field energy back to a receiver. Radar operates over great ranges and in any weather, and determines range and location simply and accurately by measuring the time required for two way energy transmission and the angle of propagation. Radar is generally most effective at frequencies higher than those normally used for radio communications.



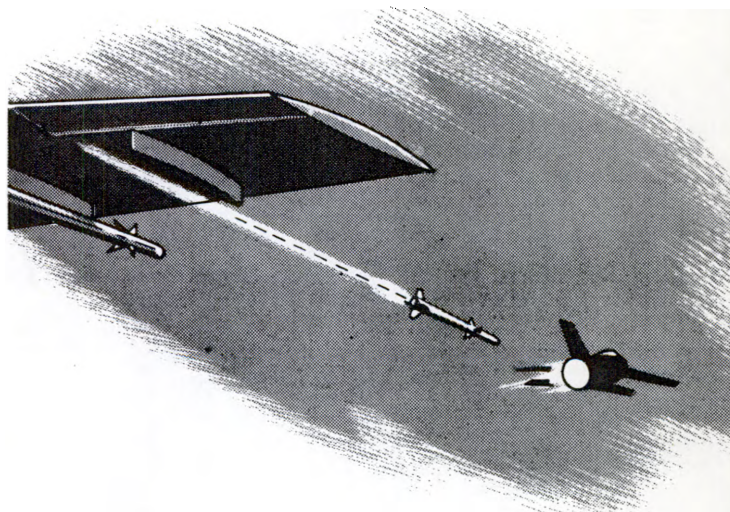
television

Television is employed in various weapons systems as a method of obtaining target information. Television cameras have been located in unmanned remote controlled aircraft or drones to televise ground targets and to send their sensed information back to control centers for evaluation. Television has also been used to remotely control missiles or bombs, permitting a more effective visual guidance system to actuate the carrier. Television has generally made use of the electromagnetic frequencies in the range between radio and radar applications for transmission. The sensor units operate on the visual band of the spectrum. Radar may be considered a form of low resolution TV.



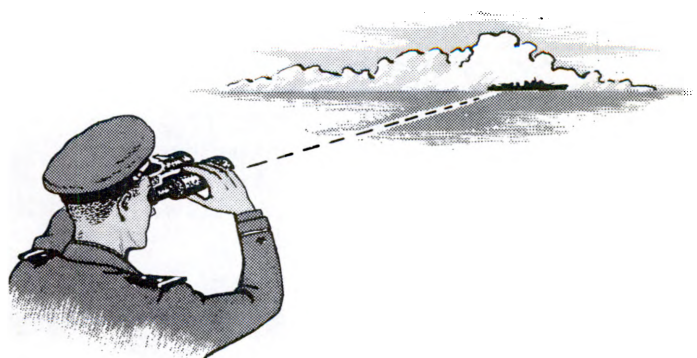
infrared

The infrared portion of the electromagnetic spectrum lies between radar and visible light. Infrared waves are known as heat waves although they do not travel through a medium in the form of heat, but are transmitted in the same manner as radio or light waves. Infrared waves are produced by thermal agitation of molecules and produce heat in any object that absorbs them. Surface molecule agitation produces electromagnetic fields that are radiated, reflected or refracted in a manner similar to light waves. The effectiveness of an infrared detection system depends on the sensitivity of the heat seeking sensor that is used to locate and position the target, and the associated circuitry that is employed by the system of tracking and for information gathering. Because of the nature of matter, and the vast expenditure necessary to create effective heat shields, there is no economical way by which an enemy can camouflage this self radiated heat. Devices which detect infrared radiation are used as sensors in heat seeking missiles that guide the missile to a target kill area.



visible light

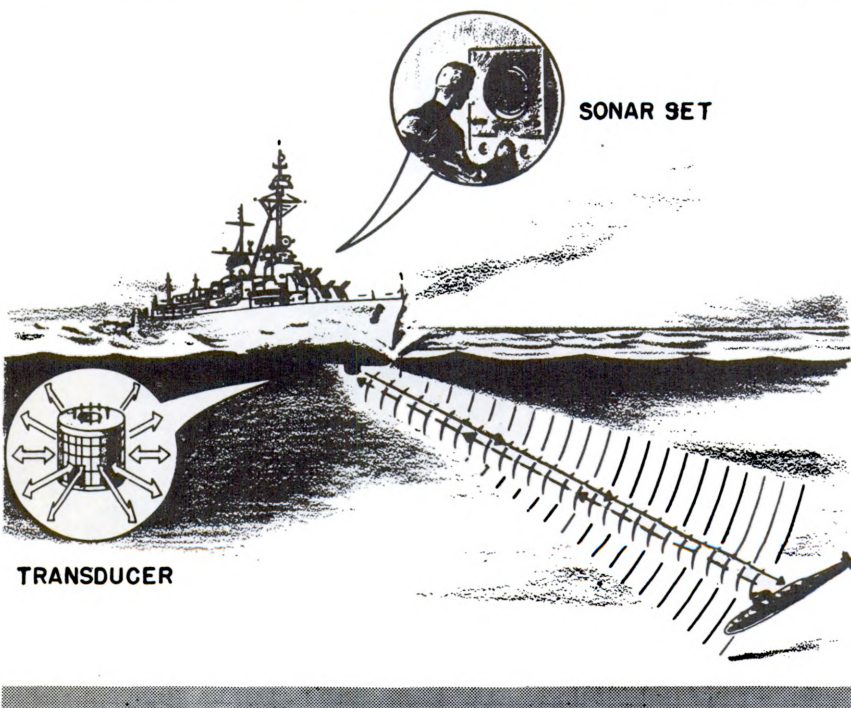
The visual end of the spectrum is still effectively used for detecting stationary low speed targets. Visual sensors include the human eye, television, and photographic equipment, and are aided by optical lens systems, telescopes and, under conditions of low visibility, by added sources of light such as searchlights and flares. Visual detection is used more often in air and to a lesser degree underwater. Underwater visual detection must be aided by illuminating sources as human perception is poor at the low light levels involved. Television has been found to be useful for underwater surveillance because of its sensitivity and because it permits remote detection.





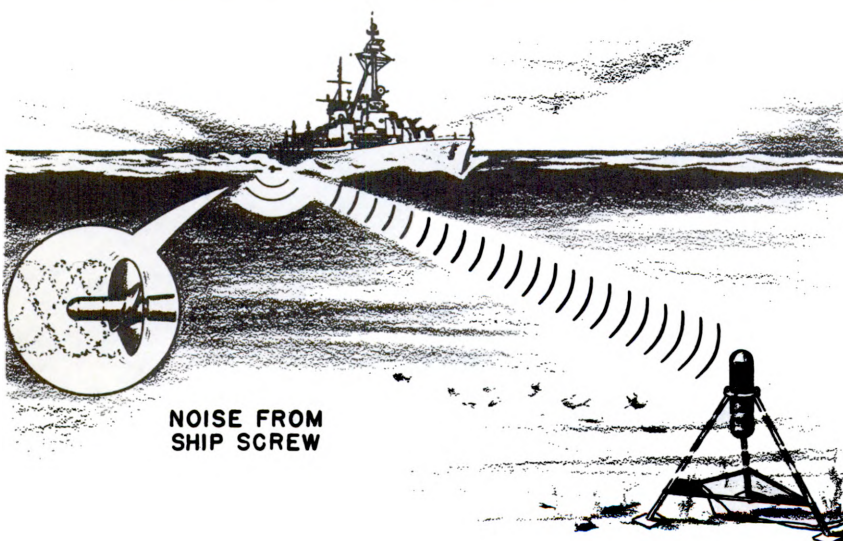
acoustics

Acoustics or sound detection systems were used in the early days of warfare for the detection, location, and tracking of aircraft. These systems made use of large horn microphones, manually operated, to detect the approach of aircraft. Similar devices have been utilized by the navy to determine the presence and position of submarines and ships. Hydrophones sense vibrations given off by underwater targets. The operation of the device is similar to the operation of a radio antenna. Wave motion in the medium induces voltages across coils in the phone unit, which after amplification can be used to track the position of a moving craft. Such systems became more and more inaccurate as increasing speed permitted the missiles to outrun the sound. The advent of radar has rendered tracking of aircraft by acoustic means obsolete.



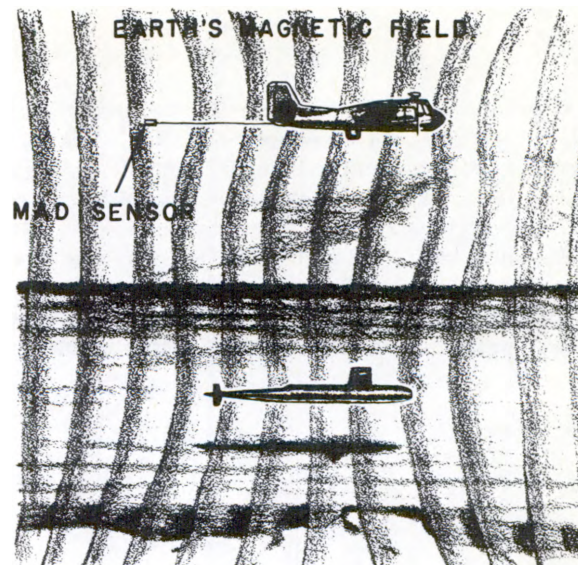
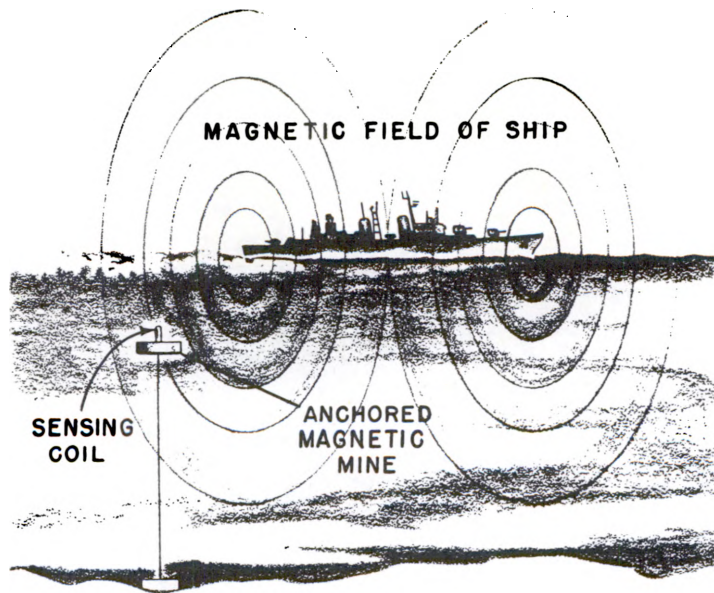
sonar

The present and major application of acoustic sensors is for use in underwater target detection systems called sonar systems. The word "Sonar" abbreviates Sound Navigation and Ranging systems and includes all types of underwater sound detection devices employed for detection, depth indication, echo ranging, and ship-to-ship underwater communication. To date, sonar has been the most effective method of undersea detection. Radio, radar, and infrared have proved ineffective because their range of transmission in sea water is practically nil.



acoustic mine sensors

Acoustical mine sensors utilize a diaphragm type mechanism which vibrates when sound waves strike it and exerts pressure on an internal crystal activating the mine.



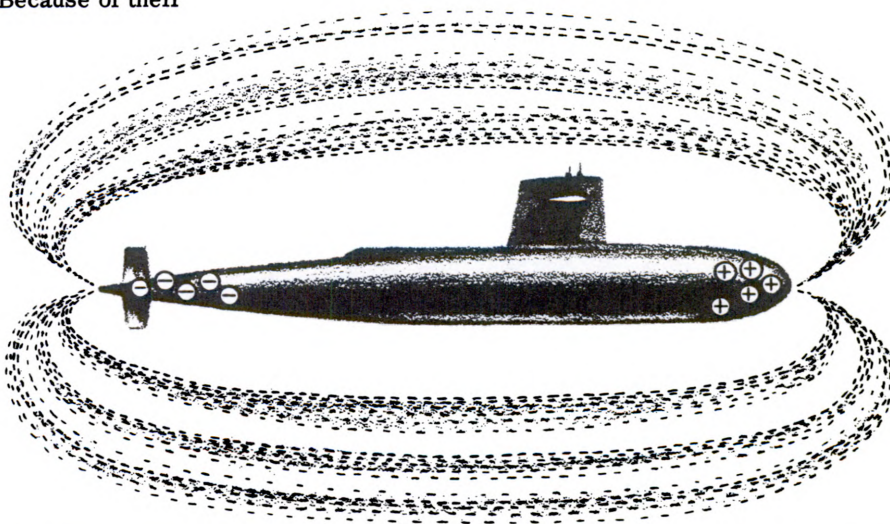
magnetostatics

Large bodies of steel, such as ships hulls and submarines, become partially magnetized as an effect of the earth's magnetic field and their own permeability and develop a magnetic character or signature of their own. Such magnetic fields can be detected by sensing coils or magnetometers and are so utilized by some target detection systems. For example, a magnetic induction mine can be set to detonate only when approached by the strong magnetic field of a ship. Because of their

size and the high permeability constant of their hulls (steel), submarines have a concentrating effect on the earth's magnetic field in their vicinity even though they are underwater. Such a variation or anomaly of the Earth's magnetic field can be detected by a magnetometer, commonly called a magnetic anomaly detector (MAD). This equipment is generally airborne and is commonly used to search for submarines.

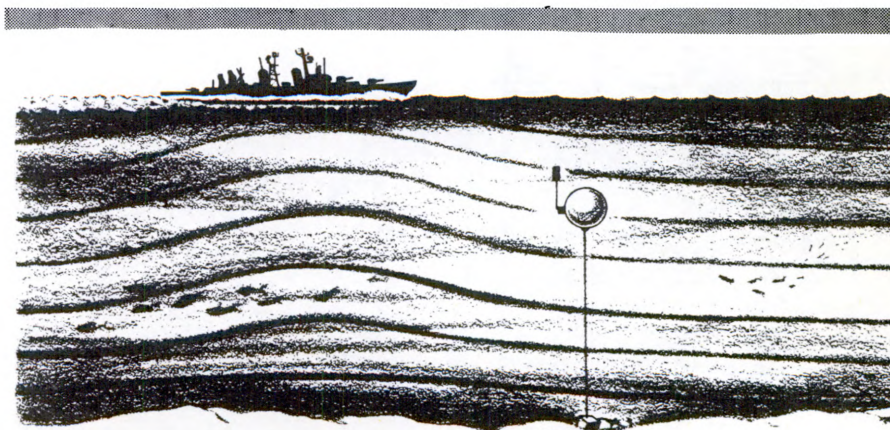
electrostatics

Metal hulls of ships and submarines develop electric potentials between various points on their structures because of capacitive effects, electrolytic action, or as a result of internal electrical systems circulating through the frame. The resulting electric field surrounding the hull can be detected at short range by suitable electrode type sensors. This phenomenon may be utilized as a proximity fusing agent in airborne and underwater missiles and mines.



hydrostatics

The motion of a ship or a submarine through the water creates a variation in water pressure detectable at quite a distance. This change in pressure, or pressure signature is utilized in some target detecting systems, particularly in mines, similar to the manner in which magnetic signatures are used.

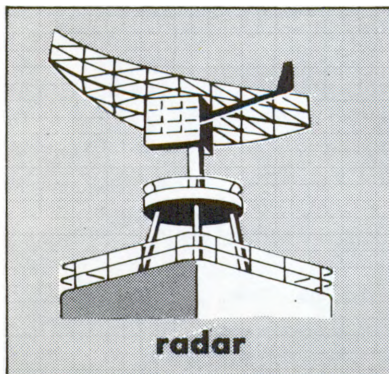


DETECTION SYSTEM

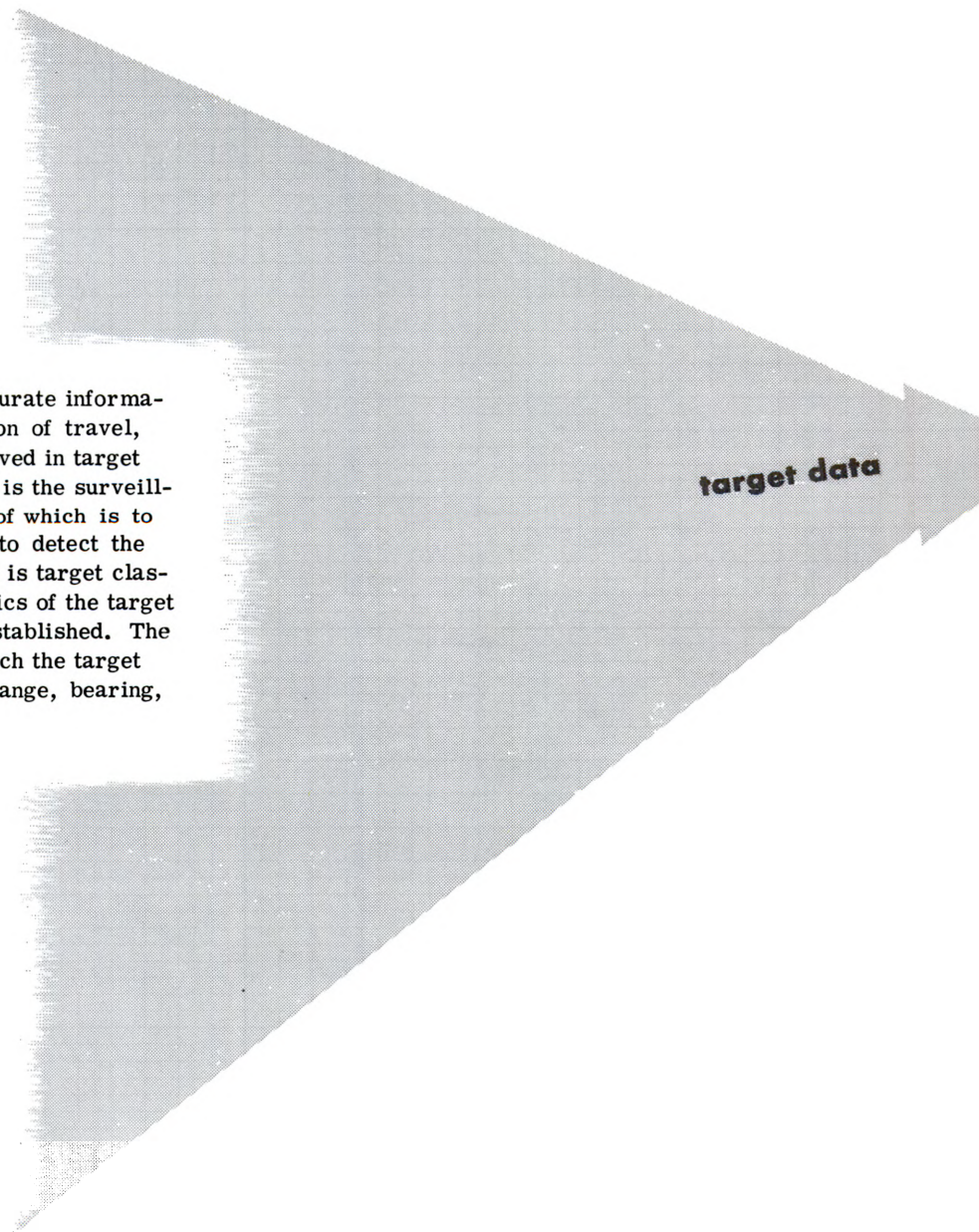
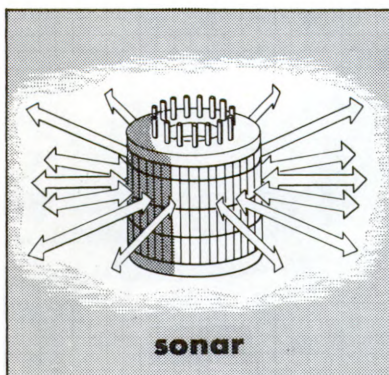
An enemy target can be either fixed or mobile, travel in the atmosphere, in space or underwater, or be cleverly concealed in camouflaged underground or undersea installations. They can be guided or unguided, maneuverable or rigidly traveling predetermined flight paths. They may vary in configuration, and may travel at speeds that range from the subsonic to the supersonic. To counteract their effects it is necessary to sense their presence, determine their capabilities, and con-

tinuously track their movements to allow an interception and destruction to their kill potential.

Detecting systems must be designed to cope with various target and environmental characteristics. Among the target characteristics to be considered are location, speed, direction, size, and type. A special consideration applies to multiple targets. A measure of a detecting system is its ability to distinguish between or resolve individual targets in a multitarget group. Once the



Once a target is sensed, precise and accurate information is required of its position, direction of travel, speed, and size. Three phases are involved in target detection in a weapon system. The first is the surveillance and detection phase, the purpose of which is to search a predetermined area for, and to detect the presence of, a target. The second phase is target classification, in which various characteristics of the target such as speed, direction, and size are established. The final phase is target location, during which the target position is localized and fixed as to its range, bearing, and depth or elevation.

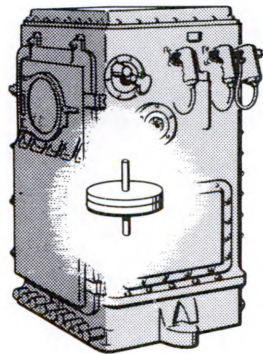
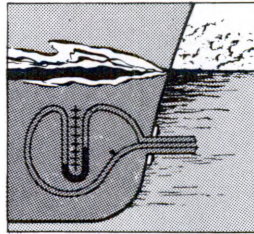


REQUIREMENTS

individual targets are resolved, the weapon system must decide how to handle them.

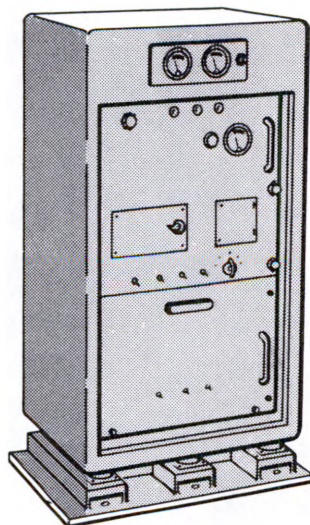
The detection of targets by any system is complicated by noise generated within the system and by ambient noise generated in the environment surrounding the target. Noise, which is any signal other than the target signal, may obscure the target. Detection systems must be designed to minimize noise and to distinguish target signals from ambient noise.

To increase the efficiency of the weapon system, it is not only necessary to detect, classify, and position the target, but is imperative to sense and gather data involving the weapon station to be utilized. When the weapon system involves a gun or launcher, sensors are required to gather data on the orientation of the weapon relative to its platform, and on the orientation of the platform relative to a frame of reference. In addition, platform characteristics and ship data (including speed and direction) must be analyzed and fed into computing networks for assimilation and correction of planned weapon trajectories.



weapon station data

environmental data



Sensors are also employed to gather information concerning the environmental factors that may affect the efficiency of the weapon system. Such environmental data includes: temperature, water and air pressure, humidity, solar radiative effects, salinity of sea water, and motion of the weapon station. In computing a missile's flight path whose terminal destination may be thousands of miles from the point of launcher, and where the flight environment may be underwater, in the atmosphere, and in a space for periods of time during its flight, environmental conditions can prove to be an extremely important factor. Although a guided missile has within its own structure a guidance system method of correcting undesired deviation, extremely powerful environmental forces (hurricanes, high wind gusts, pressures, etc.) may alter the flight path beyond redemptive measures. Thus it is imperative that measuring stations be installed at strategic locations to relay this information to a launch centers for correction correlation in flight trajectories.

METHODS OF DETECTION AND TRACKING

There are three general methods by which sensing agents are utilized:

ACTIVE

SEMI ACTIVE

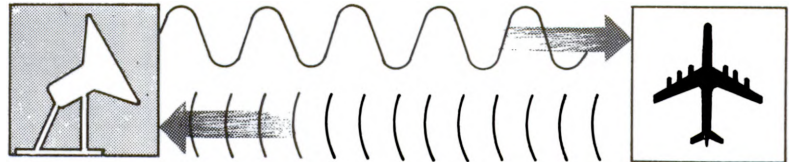
PASSIVE

DETECTION METHODS

characterize the applicable detection systems.

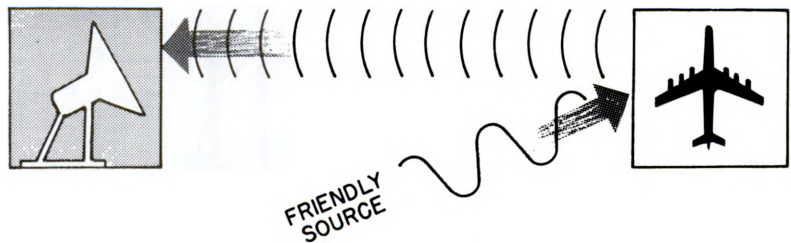
active detection

applies when the detection system itself is the source of the sensing agent. In a radar system for example, the transmitter and antenna generate and radiate the pulsed energy form which is then reflected back by the target to the radar receiver for identification.



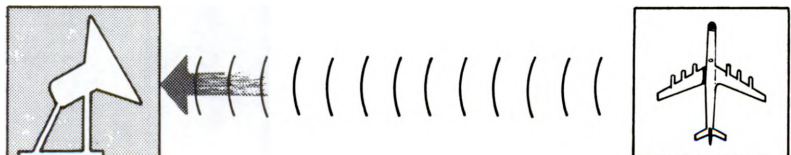
semi-active detection

refers to those systems where the sensing agent radiation is actively transmitted from a source separate from the detecting system. In some target seeking missile systems, the target is irradiated or illuminated by a radar transmitter on the ground or on the launching aircraft. The missile detecting system, which has only a receiver, then homes on the signals reflected off the target.



passive detection

methods are used when the target itself is the source of the energy signal. In such cases the detection system needs only to receive and detect the signals propagated from the target. Examples are those used in IR heat seeking missiles where the heat transmitting areas of the target emit infrared radiation, or in sonar listening systems where machinery generated noises are transmitted by the target. Some detection systems which sense the radiant energy from a natural source reflected by a target are referred to as semipassive systems.



COUNTERMEASURES

Countermeasures (CM) are any means employed by an enemy to reduce the effectiveness of a detection system. Many CM techniques have been developed including

JAMMING
DECEPTION
CONFUSION

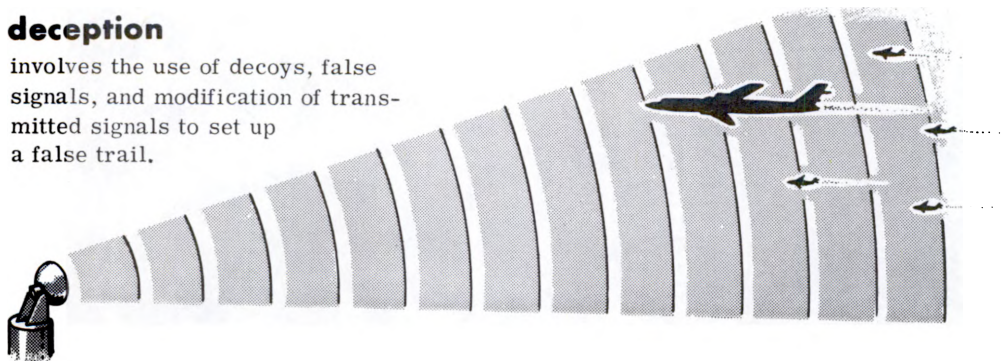
jamming

involves transmission of noise in a particular frequency band to obscure targets.



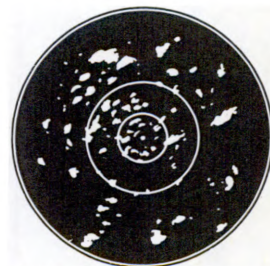
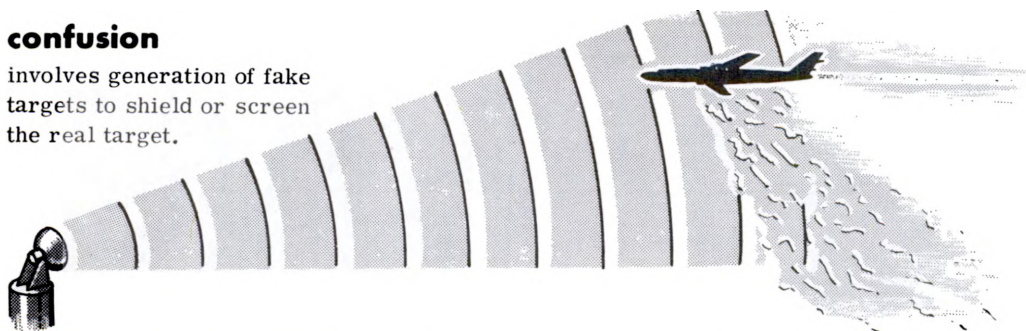
deception

involves the use of decoys, false signals, and modification of transmitted signals to set up a false trail.



confusion

involves generation of fake targets to shield or screen the real target.

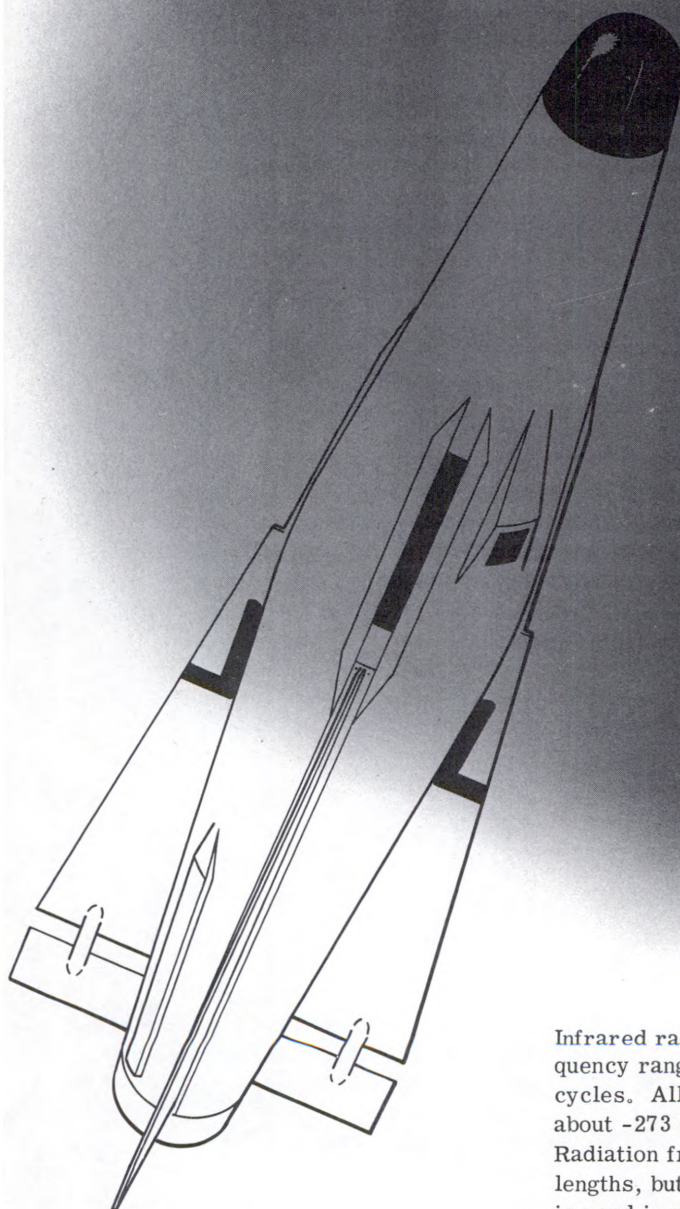


In order to operate effectively in a countermeasure environment, the detection system must employ counter-countermeasures, (CCM). CCM tactics generally involve electronic circuits and detection techniques which will counteract or bypass the effects of the enemy CM tactics.

Many overall detecting systems utilize multiple sensing devices in a single installation to increase the probability of target detection, and as a means of overcoming countermeasure techniques. In some cases the same type of sensing equipment is used but with varying

characteristics. For example, multiple radar sets operating at different frequencies make jamming and distortion practically impossible. In other cases one operation may utilize different kinds of sensors; for example, an ASW aircraft (helicopter) might utilize sonar, MAD, and IR equipment concurrently to detect a submarine by proper correlation of the data obtained from the different sensors. Another multisensor application is a missile that utilizes one system for initial guidance, another for midcourse guidance, and a third for terminal guidance and proximity detonation.

INFRARED

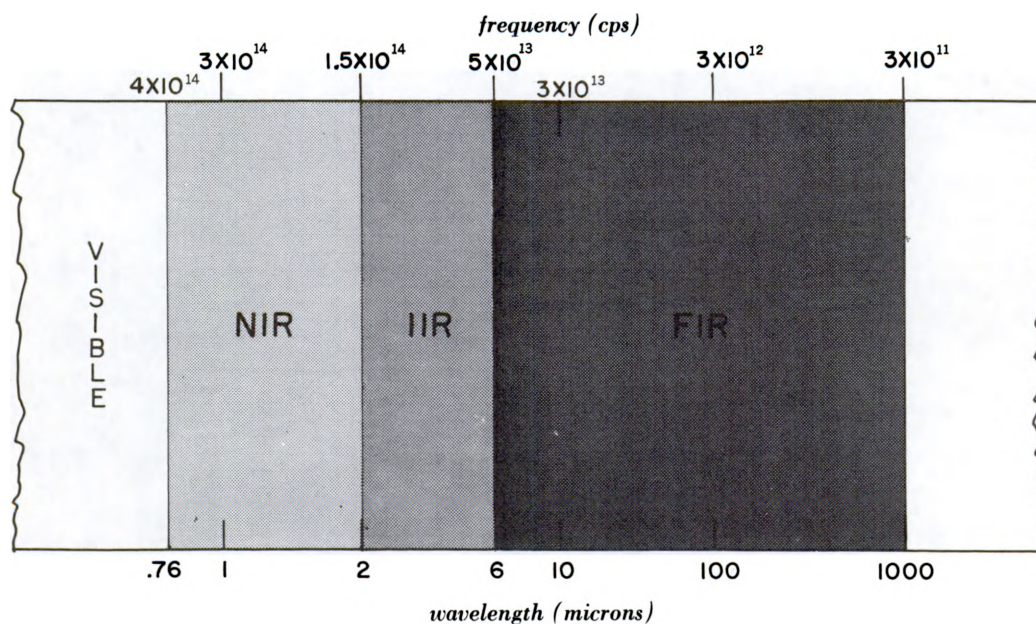


Infrared radiation is a form of electromagnetic energy. Its frequency ranges from approximately 1 million to 500 million megacycles. All objects above absolute zero (zero degrees Kelvin, about -273 degrees C) emit a degree of infrared radiation. Radiation from any solid is emitted over a wide range of wavelengths, but it peaks at one particular wavelength, a fact which is used in military applications. Detection of infrared energy depends on the contrast between IR from the target and the radiation emitted by the background. A cold object with a warm background is just as good a target as a warm object with a cold background.

infrared spectrum

The interval of the spectrum between visible light and the microwave region used for high definition radar is known as the infrared or radiant spectrum. The infrared portion of the spectrum starts at the red end of the visible spectrum, and extends to the millimeter wave portion of the radio segment. Because the frequencies in the infrared spectrum are in the millions of megacycles, it is customary to refer to wavelength rather than frequency when describing the specification of these waves. The unit of wavelength most commonly used is the micron (μ), which is 10^{-4} cm or 10^{-6} meters in length. In microns, the infrared spectrum extends from 0.76 micron to 1000 microns.

For practical purposes it is convenient to think of the infrared spectrum as comprising three broad regions; the near-infrared (NIR), the intermediate-infrared (IIR), and the far-infrared (FIR). The NIR region is from 0.76 microns to 2.0 microns and is used primarily for communication purposes. The IIR region is from 2.00 microns to 6.0 microns, and is used for the passive detection of targets with a relatively high emissive output (aircraft). The FIR region is from 6.0 to 1000 microns, and is utilized for the detection of targets with a low emissive output (ships wake, etc). Active detection devices utilize primarily the IIR region for their range of operations.



methods of detection

There are two commonly used methods of detecting a target by means of infrared radiation. Characteristic radiation infrared detection is the term used to describe the passive process of detecting emitted infrared radiation which is characteristic of the target. The second method operates on the principle that infrared radiation is detected when a beam of infrared is radiated by a special transmitter, and then reflected by the target. A third method, semiactive, is not generally used. It depends on the reflection of infrared energy from a source independent of both the sensor and target.

The passive method of infrared detection has distinct advantages over the active method because the former permits the use of a simple detecting device and does not need a transmitter.

The detection of infrared radiation, no matter what the method employed, depends on the principle of measuring a change in the physical characteristics of the transmitting media resulting from the emitted or reflected radiation. Sensitive thermal detectors are utilized for the detection and surveillance of radiating sources, and have the advantage of being less susceptible to counter detection and interference than radar or visible light.

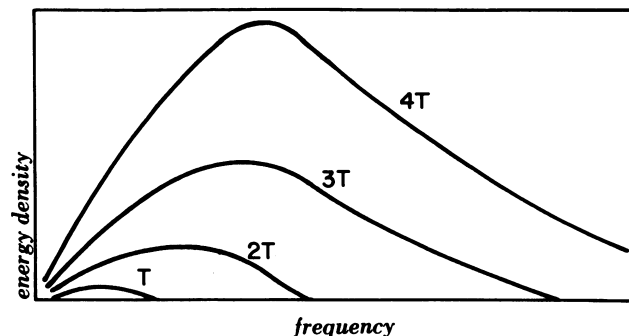
BASIC PHYSICAL LAWS

Specific physical relationships or laws govern the design and operation of all IR detecting equipment. The basic laws which describe the characteristics of IR are developed first for black body radiation, the theoretically ideal case, and then modified to describe the radiation from any other source.

black body

The visible color of an opaque body depends on the frequency of the light reflected by that body. If none of the light is reflected, i.e., all the radiant energy that strikes it is absorbed, then the body would be an ideal black body. Its reflectivity would be zero and its absorption constant would be 100%. A black body is defined as an object that is both a perfect absorber and a perfect radiator of all radiant energy that is a function only of temperature. Such a black body then acts as a perfect source of IR, radiating at any given temperature with the greatest and most uniform intensity possible. The closest approximation to an ideal black body is an almost completely closed cavity in an opaque body. The simplest type is usually an elongated, hollow metal cylinder, blackened inside, and completely closed except for a narrow aperture at one end. Light or other radiation entering the opening is almost completely trapped by multiple reflections from the walls, so that the opening usually appears completely black. The total emission of radiant energy from a black body takes place at a rate expressed by the Stefan-Boltzman (fourth-power) law, while its spectral energy distribution is described by Planck's radiation formula, or Wien's laws.

Experimental study of the spectral distribution of the intensity of black body radiation yields two important facts. One is that the total intensity, integrated over all wavelengths, is proportional to the fourth power of the absolute temperature, and second, that the maximum value of intensity moves toward high frequencies at higher temperatures, in accordance with Wien's displacement theory, which will be discussed later in the chapter.



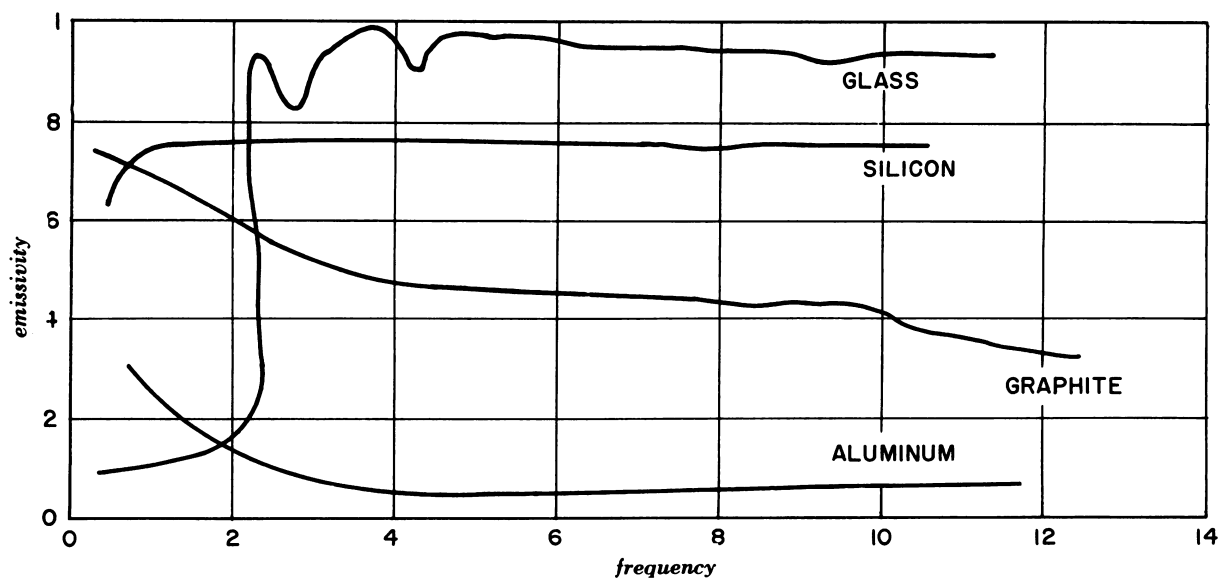
emissivity

To compare the radiation emitted by an actual source with that of a black body, the concept of emissivity must be defined. Emissivity is the ratio of the total radiant energy emitted by an object at temperature T , to the total energy emitted by an ideal black body at the same temperature. The emissivity of a perfect radiator or black body is considered to be unity; therefore, for all other objects emissivity must be less than one. All bodies emit radiation as a result of the thermal agitation of their molecules or atoms, whether there are other causes of excitation or not. The degree of radiation depends upon the nature of the body and upon its temperature. The emissive characteristic of a surface at any temperature is the rate at which it emits energy of all wavelengths per unit area of radiating surface. The emissivity of any object depends on its ability to absorb energy. If the surface absorbs a large percentage of the total infrared energy striking it, then its emissivity ratio will be quite high. If the surface reflects most of the incident radiation, the object will have a low emissivity characteristic. The emissivity of various surfaces is shown in the following table.

emissivity of various surfaces

Surface	Emissivity
Black Body	1.00
Lamp Black	0.95
Steel	0.60
Aluminum Paint	0.25
Steel Stainless	0.09
Aluminum For Aircraft Structure	0.08
Aluminum Foil	0.04
Silvered Mirror	0.02

Emissivity of most surfaces varies with wavelength.
The relationship of emissivity to frequency for various materials are as illustrated.

**TOTAL EMISSIVE POWER**

The total emissive power of an ideal black body is expressed by the Stefan-Boltzman law and is stated by the formula:

$$E = \sigma T^4,$$

where T is the absolute temperature ($^{\circ}\text{K}$)

E is the radiant energy in watts cm^2
per hemisphere

σ is the Stefan-Boltzman constant (5.7×10^{-12}
watts $\text{cm}^2 \text{ } ^{\circ}\text{K}^{-4}$)

When the black body is radiating at a temperature other than absolute zero, the law may be written as

$$E = \sigma (T_4 - T_a^4)$$

where T_a = absolute temperature of the surrounding air.

When a surface is not an ideal radiator, the difference in emissivity is accounted for by an emissivity factor and the law is expressed as:

$$E = \rho \sigma (T^4 - T_a^4)$$

where ρ = emissivity factor.

From the previously stated relationship it is easily determined that if the temperature of a surface is doubled, the radiation emitted from the object will be increased 16 times.

SPECTRAL DISTRIBUTION OF POWER

The total emissive power of a surface is a function of both wavelength and temperature. From a study of the spectral energy distribution of a thermally excited black body, a German physicist, Wilhelm Wein, formulated the Wein laws which show a correlation between peak frequency and temperature, and the relationship between wavelengths at which maximum radiation occurs and activating temperature. The following three laws relate to the radiation relationships that exist in a black body.

1) The wavelength λ_m , for which the radiation achieves the greatest intensity, is inversely proportional to the absolute temperature of the black body.

$$\lambda_m = \frac{\sigma}{T}$$

where λ_m is the wavelength of the maximum radiation
 T is the temperature of the body °K
 σ is the constant for a given surface.

As the temperature rises, the peak of the distribution curve is displaced or shifted toward the short wavelength end of the spectrum. This relationship is expressed as Wien's displacement law. The value of σ for a black body is about 0.2897 centimeter-degree; the value of σ for platinum is 2.630. This formula only locates the point of maximum radiation, and not the radiated energy in other parts of the spectrum.

2) The emissive power of the black body within the maximum intensity wavelength interval ($d\lambda$) is proportional to the fifth power of the absolute temperature:

$$dE_m = CT^5 d\lambda$$

where C has been shown to be 3.732×10^{-12} watts per cm^2 . This will be developed later as an explanation of Planck's Distribution law.

3) The spectral energy distribution of radiation from a black body at temperature T can be expressed as:

$$dE_\lambda = C_1 \lambda^{-5} e^{-C_2/\lambda T} d\lambda$$

where dE_λ is the emissive power within the wavelength interval $d\lambda$

C_1 and C_2 are constants

This law is identical with Planck's equations at short wavelengths, but fails at longer wavelengths. Planck's distribution law for the emission of thermal radiation within an interval $d\lambda$ from a unit area of a black body into a hemisphere can be expressed as:

$$dE_\lambda = \frac{C_1}{\lambda^5} \frac{1}{(e^{C_2/\lambda T} - 1)}$$

where dE_λ = the emissive power of unit area in wavelength interval

$$C_1 = 3.732 \times 10^{-12} \text{ watts cm}^2$$

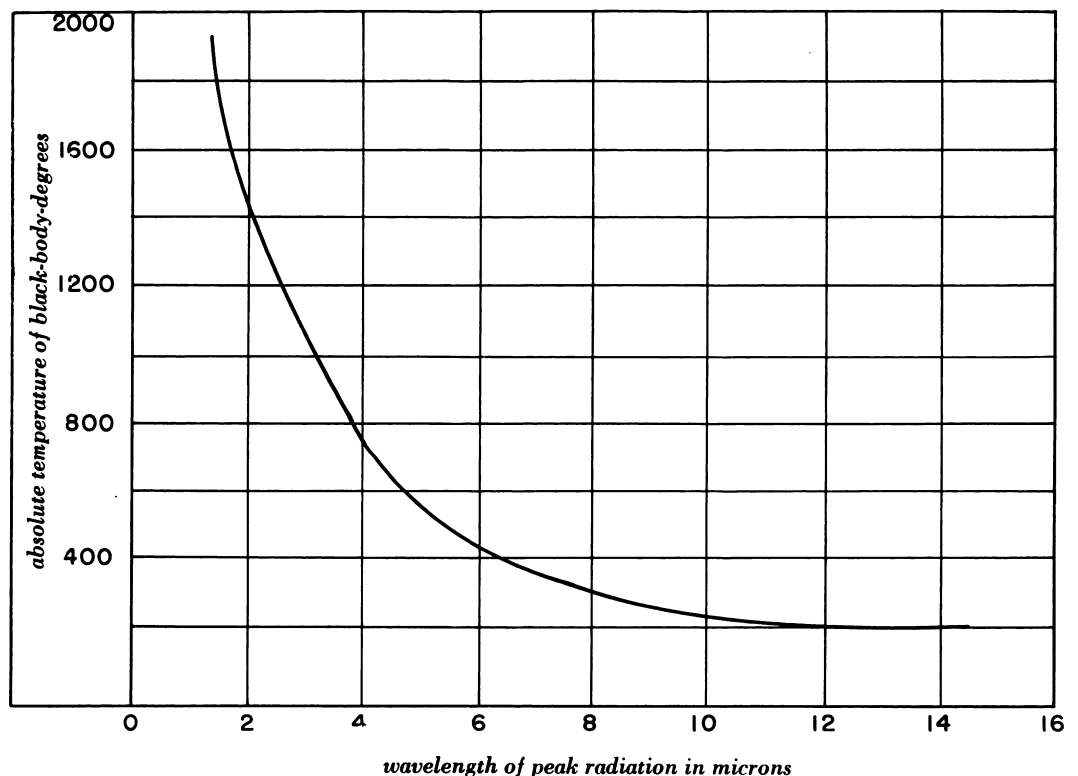
$$C_2 = 1.436 \text{ cm (}^\circ\text{K)}$$

$$\lambda = \text{wavelength in cm}$$

$$T = \text{absolute temperature (}^\circ\text{K)}$$

From the preceding development it can be shown that although IR from a source will be distributed over a large part of the spectrum, maximum radiation occurs at a specific wavelength. It can also be developed that peak energy output varies with temperature, as does the wavelength at which the peak occurs.

the wavelength of the peak radiation from a black body in relation to its temperature



TRANSMISSION CHARACTERISTICS

Infrared radiation is a form of electromagnetic radiation, and is transmitted in the same manner that a radio or light wave is transmitted. Infrared waves are known as heat waves because they are produced by thermal agitation of the molecular structure of a radiant surface, producing a degree of heat emissivity from that surface. Infrared waves do not travel through space in the form of heat, but are propagated as electromagnetic waves; the heating of the atmosphere is negligible when infrared radiation passes through it. Environmental conditions that exist in the medium being traversed seriously attenuate in the infrared propagation. Rain, fog, snow, and water attenuate infrared in a manner similar to light wave attenuation. Like light, infrared is reflected, refracted, and subject to atmosphere attenuation, thus requiring line of sight for transmission and reception. The range of infrared equipment is usually expressed in average clear weather (ACW) distances.

atmospheric transmission

In military applications of IR systems, the transmitting medium is generally the atmosphere. The effect of the atmosphere on the transmission of IR is a very important factor in determining the overall effectiveness of the system. There are two primary causes of atmospheric attenuation: scattering and absorption. These two influences are additive, but absorption is usually the more important consideration.

SCATTERING

Generally, scattering denotes the change in direction of electromagnetic radiation resulting from the inhomogeneity of the transmitting medium. When IR enters the atmosphere part of it is diffusely reflected, or scattered, in all directions. This is caused by the interposition in the medium of particles of varying size, from microscopic specks to electrons, and the deflection from the encounter with these small bits of matter. Rayleigh (an English physicist) determined that the intensity of the energy of

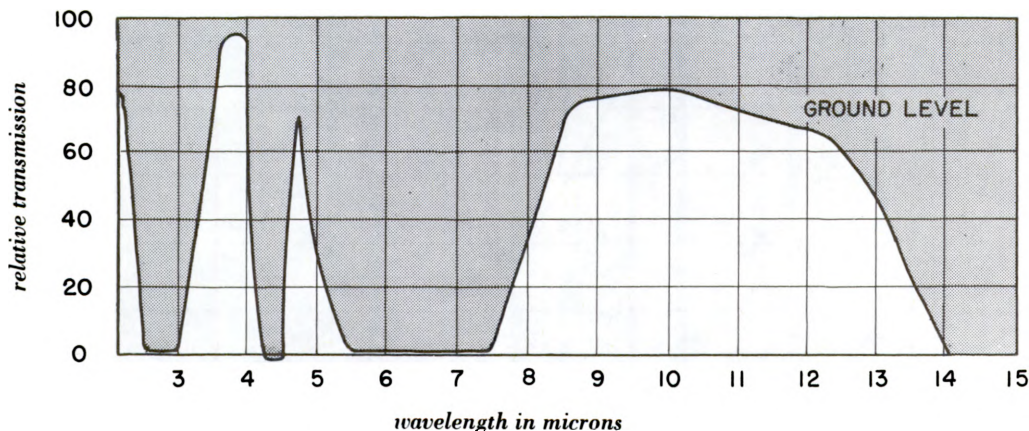
wavelength λ scattered in any direction making an angle θ with the incident direction is directly proportional to $1 + \cos^2 \theta$ and inversely proportional to λ^4 . From this determination it follows that the scattering effect increases with frequency or is greater at shorter wavelengths, and often at the shorter infrared wavelengths scattering predominates over absorption.

ATMOSPHERIC ABSORPTION

Absorption is the process whereby a portion of the infrared energy is transferred to the medium substance that is being traversed. A measure of the rate of decrease in intensity of the IR beam in its passage through a medium is called the attenuator constant of the medium. When IR radiation enters a body of matter, it experiences two types of attenuation. A portion of the energy is subjected to scattering, while another portion is absorbed or converted into another energy form. The absorbed portion ceases to exist as radiation, or may be re-emitted as secondary radiation. The absorption constant must be thought of as separate from, but related to, the scattering coefficient. In fact the total effect is usually referred to as the attenuation constant of the medium.

The chief cause of IR absorption in the atmosphere is the water vapor and carbon dioxide content of the medium. The molecular structure of these substances allow them to combine chemically or absorb a portion of the IR energy at particular frequency levels. A nonmathematical explanation is that the combination takes place when the molecular structure of the absorbant is resonant or receptive to the action. The frequency bands that are most receptive to this interaction are known as absorption bands, and the portions of the spectrum between these bands (or the nonabsorbant regions) are known as transmission bands or windows. Windows exist in the atmosphere in the following wavebands:

1.2 - 1.3 microns	3.0 - 4.00 microns
1.5 - 1.8 microns	8.0 - 14.00 microns
2.0 - 2.4 microns	



ATMOSPHERIC TRANSMISSION CHARACTERISTICS

The operational usefulness of an IR detector depends on its ability to detect or be responsive to wavelengths where windows exist in the atmosphere. The body temperature of a human produces a peak radiation near 9 microns, and it would be necessary to utilize a detector capable of operation in the 8- to 14- micron window area for target location. As no equipment at the present time is completely operational in this range, humans must be illuminated with radiation from a NIR source, if detection is to be successful. Another aspect of absorption and its effect on transmission of IR is the density of the medium being traversed. At low altitudes where the atmosphere is denser, absorption increases and transmission efficiency decreases. This is true regardless of the medium; an increase in density means an increase in absorption.

transmission through liquids and solids

Accurate measurements of electromagnetic type radiation which have traversed various thicknesses of matter have established that each infinitesimal layer perpendicular to the direction of propagation decreases the intensity of the radiation by an amount which can be expressed by the exponential expression when the medium has been penetrated an amount x :

$$I = I_0 e^{-ax}$$

I_0 is the flux intensity at entrance into the medium;
 a is the absorption coefficient.

The absorption coefficient divided by the density of the absorption medium is called the mass absorption coefficient. The transmission of infrared radiation through liquids and solids depends on the absorption and reflection coefficients at the surfaces of these materials. The intensity of the transmitted radiation can be found from the expression

$$I = I_0 - I_{\text{ref}} - I_{\text{sc}} - I_{\text{abs}}$$

where I_0 = flux intensity at entrance

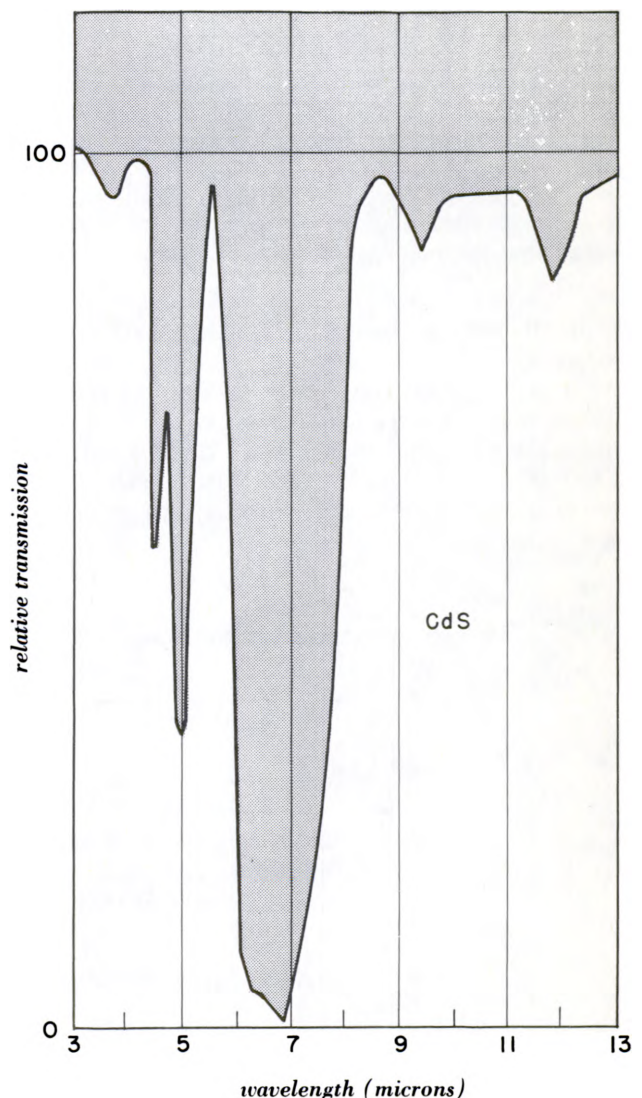
I_{ref} = loss resulting from reflection

I_{sc} = loss resulting from scattering

I_{abs} = loss resulting from absorption

Many liquids and solids are employed in the optical systems of IR detector devices as filters or lenses, and although the lengths of the transmission paths in these systems are insignificant compared to ranges between target and detector, their absorption effects because of greater densities are significant. Absorption is generally the limiting factor for IR transmission, as is evident from the lack of windows which are characteristic of many liquids and solids. The bands illustrated for carbon disulfide are typical for most liquids.

For the NIR region, glass and quartz are satisfactory. They do however, reduce the 3- to 4- micron sensitivity of the lead-sulfide cells normally used in this region. The loss in sensitivity is a function of the absorption loss of the optical material used. For the FIR, single crystals of silver chloride, rolled flat, are satisfactory windows for the transmission of the far infrared. Sodium chloride (rock salt) grown in single crystals, cut and ground into a lens or window is excellent but hygroscopic. A mixture of thallium-iodine and bromide, a German developed crystal, transmits up to 30 microns. For transparent synthetic or natural crystals, absorption is relatively less serious, and new developments are allowing satisfactory transmission characteristics over the entire windowed ranges.



ABSORPTION IN LIQUID

INFRARED SENSORS AND DETECTORS

IR detectors and sensors operate in accordance with physical and chemical laws associated with thermal measurements of various kinds, photoelectric effects, luminescent manifestations, and chemical reactions. In

addition to the human eye, which is sensitive to only a small range of IR (spectral range of 0.3 to 1.0 micron), there are 4 principal types of IR detectors and sensors:

1. THERMAL TYPE DETECTORS
2. PHOTODETECTORS
3. LUMINESCENT DETECTORS
4. CHEMICAL DETECTORS

THERMAL DETECTORS

The emissivity of a surface, or its thermal radiation characteristic, depends upon the structure of its body and its temperature. Thermal radiation is observed and measured by various types of heat sensing devices. The heat sensor types are normally categorized as follows:

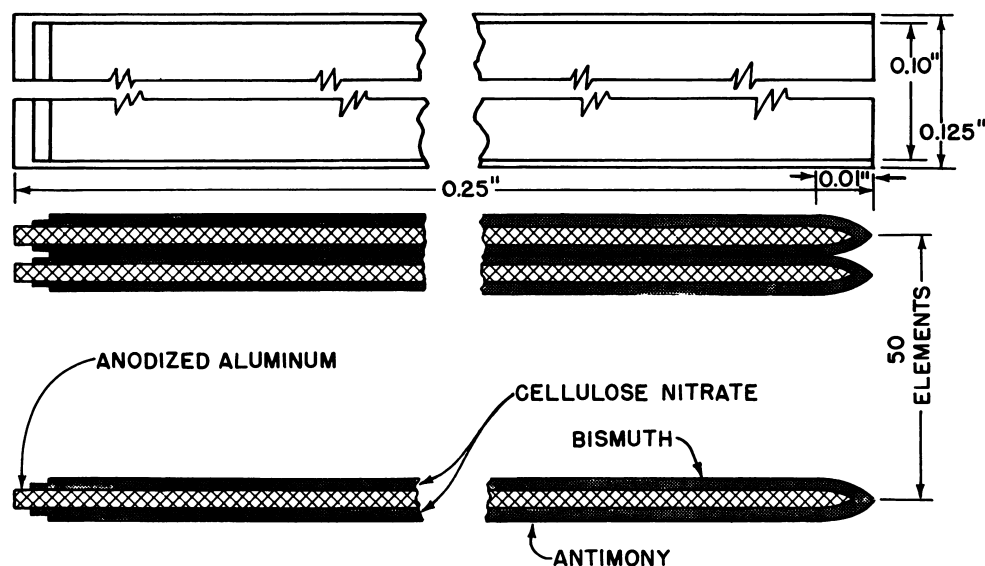
1. Radiometers which are instruments for detecting and measuring radiant energy (by absorption),
2. Thermocouples and Thermopiles which are devices for measuring generated voltages caused by temperature differentials in two dissimilar metals as a result of heat exposure,
3. Thermometers and bolometers which are sensitive instruments utilizing the principle of change in resistance characteristic of particular substances when exposed to IR, and
4. Heat cells which measure pressure-volume changes in gas when its temperature is increased or decreased.

RADIOMETERS are extremely sensitive instruments for measuring the intensity of IR radiation, both direct and scattered by the atmosphere. Among the many types of radiometers are the Crookes and the Nichols (sometimes called "bane" radiometers). Although extremely sensitive, their results are usually more theoretical than practical, and they are not normally used in military applications.

THERMOCOUPLES are devices consisting of two junctions or joints between two dissimilar type metals, with one of the junctions left at a fixed temperature. When the temperature of the second junction is influenced by a temperature change, a thermoelectric force is generated. Various combinations of metals are used, for example, antimony and bismuth; copper and iron; or copper and constantan. Often one of the junctions is enclosed or protected against temperature changes, and the other exposed to conditions to be measured.

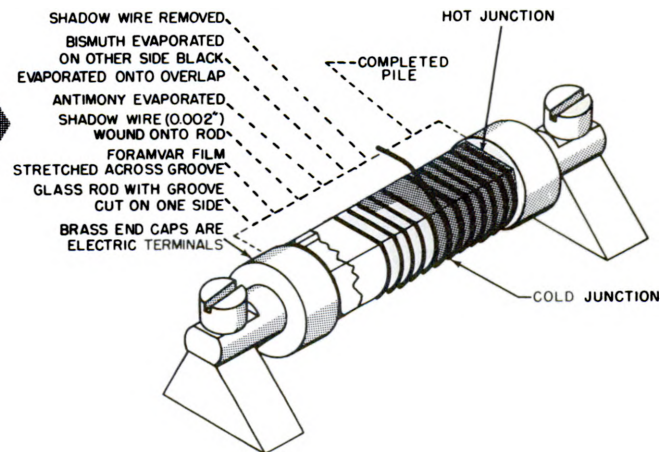
A thermopile consists of a number of thermocouples in series. Radiation, incident upon and absorbed by one set of junctions, is amplified by the series action of the thermopile and the increased electromagnetic force developed is more easily measured. By making the device rotatable, the source of highest intensity may be located. A method of thermopile construction is illustrated below.

The thermocouple elements are joined by folding a thin cellulose-nitrate film over an anodized aluminum form. One side of the film is then coated with bismuth and the other side with antimony. The film is purposely mounted to form an insulated loop at the junction point of the two metallic coatings to manufacture a thermocouple. The aluminum forms are connected in series to form a series thermopile. The junctions, backed only with a thin film, are painted black to increase the absorption rate. This method of construction permits the radiation to produce the maximum rise in temperature at the junctions, while the aluminum backing strip affords radiation for the cool junctions.



construction details of folded thermopile elements

Another method of thermopile construction is the spiral thermopile consisting of a series of thermocouples composed of a hot junction deposited on formvar film wrapped around a glass rod. An air space grooved in the glass underlies the hot junction. The conductivity of the glass keeps the cold junction cool.

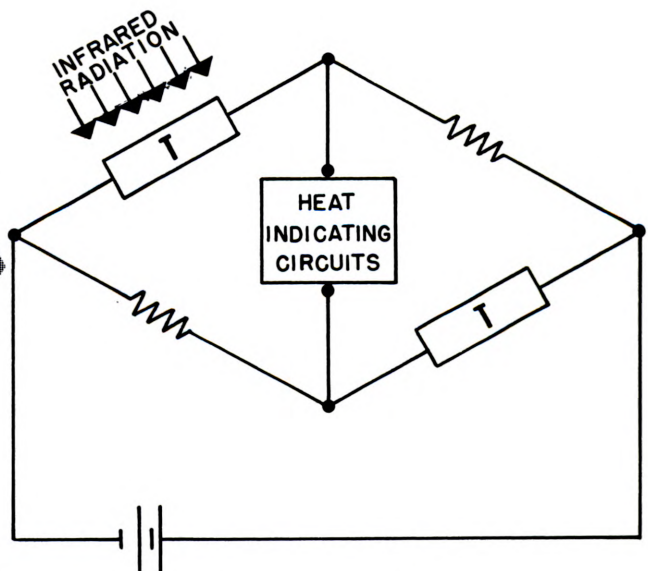


schematic diagram of step-by-step construction of a spiral thermopile

The **BOLOMETER** is an extremely sensitive thermometric instrument which depends on the change of electrical resistance of a material for its indication. It normally consists of a slender strip of platinum or a semiconductor having a high (negative) coefficient of resistance, mounted in a bridge type network. A slight amount of IR energy striking the strip causes a measurable deflection in a sensitive galvanometer in the resistance bridge coupled to it.

Two thermistors (thermally sensitive resistors) can be used in this type of detector. When FIR radiation falls on this type of absorptive material, its temperature is increased and the thermistor increases in resistance. The other thermistor, being shielded from the radiation, provides a compensating circuit for changes in ambient temperatures.

A typical IR detection system, utilizing a bolometer as a sensor for guided missile applications, would operate as follows: A scanning device (usually a mirror) revolves through a predetermined search area. When radiation from the target (reflected or emitted) strikes the mirror surface, it is immediately reflected to the heat sensing volometer for study. A change in temperature will cause an imbalance to exist in the bridge circuit; the indicator (galvanometer or other electrical measuring device) can now be used as a guidance device to seek out the heat source.



simplified schematic of a thermistor bolometer

HEAT CELLS

The Golay heat cell makes use of the principle that the pressure-volume characteristic of a gas is varied with changes in temperature. The gas is contained in a metal chamber which contains a mirror membrane whose position is affected by the pressure exerted on it. When radiant heat falls on the receiver, the temperature of the gas rises, and the accompanying increase in pressure causes the membrane mirror to move or distort.

A system of lens can then be utilized in conjunction with photocells to measure the degree of pressure change. The Golay detector has a very rapid response compared with other FIR detectors, but operates only on intermittent signals. An advantage of thermal detectors is that they are responsive to IR over a large portion of the IR spectrum; they operate effectively up to 14 microns. Thermal detection, however, requires more incident radiation than other FIR detectors to produce an equivalent output.

photodetectors

Photodetectors use photon energy of IR to vary an electrical property of the sensor material. They may be divided into three major types - photoconductive, photovoltaic, and photoemissive. The photoconductive cell utilizes an element whose resistance varies when exposed to radiant energy. The elements usually employed are semiconductors, such as lead sulfide, lead selenide, lead telluride, and germanum. They are used as radiation detectors, particularly in the NIR (10^{10} microns). The photovoltaic cell is used chiefly in photographic light meters. It produces small voltages in low resistance devices when exposed to IR, but is normally not sensitive enough for communication purposes. Photo-

emissive cells produce an electrical charge when exposed to light waves. The photon energy changes the resistance of the sensor device and produces a voltage potential which causes an emission of electrons where intensity can be determined. They are relatively insensitive but provide high fidelity and low signal-noise levels. The inner surface of the cathode element of the cell is coated with cesium oxide or another light sensitive electron emitting material. The electrons emitted are collected by the anode and then amplified for better indicating purposes. Photocells can be any of the three types discussed. The signal-to-noise ratio of the photodetector is the limiting factor in determining its effectiveness.

PERFORMANCE OF INFRARED DETECTORS				
Detector	Range (Microns)	Threshold \pm (Watts)	Time Constant (Seconds)	Scanning Rate \pm CPS
Thermocouple	Entire spectrum	10^{-10} - 10^{-11}	0.1-0.03	
Bolometer	Entire spectrum	10^{-8} - 10^{-12}	0.2-0.002	10 - 50
Phototube	0.2-1.3	10^{-13} - 10^{-16}	10^{-9}	10^4
Heat cells, Golay	Entire spectrum	10^{-10}	0.005-0.0003	
Photoconducting cells:				
PbS	0.5-2.8	10^{-11} - 10^{-13}	4 - 10×10^{-4} , 20C 2 - 7×10^{-4} , -190C	10^4
PbSe	1.0-4.5		$3 - 10 \times 10^{-5}$	
PbTe	0.5-5.5		$1 - 5 \times 10^{-4}$, -190C	
Approximate useful range				
\pm S/N = 1				
\pm Typical practicable rates of a scanning beam across surface.				

luminescent detectors

Fluorescence and phosphorescence are luminescent effects that appear as visible glows on films or screens that have been exposed to radiation. The term fluorescence denotes the process of emission of electromagnetic radiation as the result of the absorption of energy by the fluorescent system. Fluorescent materials glow only as long as the radiation continues, while phosphorescence continues to glow for some time after excitation has ceased.

chemical detectors

Photographic emulsion plates chemically coated with substances that are sensitive to shorter wavelengths are being developed by the Navy. Most of the work in this field is still in the experimental stages and a discussion at this time would be in the realm of hypothesis.

detector characteristics

The parameters used to measure the efficiency of a detector system and to be used as a basis of comparison between systems are:

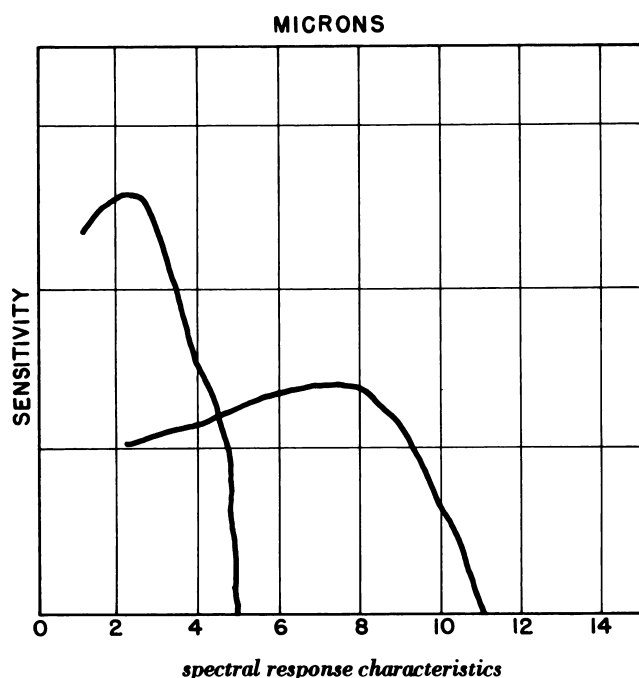
1. Sensitivity
2. Spectral response
3. Response Rate
4. Signal-to-noise ratio
5. Minimum Power Requirements for detection

sensitivity

The sensitivity of a sensor is expressed as a ratio of output power to an input radiation level. The amount of signal produced per unit of input radiation intensity is called the sensitivity of the system and is generally expressed in volts per watt. It is also important in rating the sensitivity of various systems to specify the temperature ranges, frequency of radiation, and response rate of the detector so that comparisons between sensors can be made fairly.

spectral response

An important factor in determining the sensitivity of a detector is the variation in output signal intensity with changes in the wavelengths of received radiation. When the output value of a spectral band falls to half its maximum value, its useful limit is reached. Different detectors respond to different wavelengths, and it is necessary to specify the desired area of operation before a spectral response characteristic can be determined.



response rate

An important factor in sensor efficiency is its ability to detect and respond to the sensed radiation within a specified time interval or before the signal intensity changes. The maximum scanning rate of the system is dependent on this characteristic and determines the military cases of the device. A time constant is defined as the time necessary for the sensor to reach 63% of its maximum output signal. Guided missiles that are controlled by a heat seeking device require an almost instantaneous flow of information to its guidance system to insure the probability of a kill.

signal to noise ratio

In all systems of transmission or reception of electromagnetic radiation, the undesired factor of noise must be considered. Noise can be thought of as the lower limit of amplitude that must be overcome to insure fidelity of reception. Noise sources generated in IR systems are caused by thermal fluctuations in the molecular structure of the surface, or in the electronic circuitry associated with the system. The noise equivalent power (NEP) of a detector is the input power in watts necessary to negate the effects of the noise, or to produce a signal to the noise output over the reference bandwidth. The detector with the lower NEP has the higher useful sensitivity. A sensor that has the ability to detect a target whose radiation level after reception is slightly higher than the noise level is considered extremely sensitive.

minimum detectable power

The absolute minimum detectable power (MDP) for a signal-to-noise ratio is $NEP = MDP$. This can be expressed as $MDP = NF \times C^{-T-B}$

where NF = the noise factor

C = an empirical constant

T = temperature

B = frequency bandwidth of receiver

The need for cooling to decrease the noise level is fairly obvious.

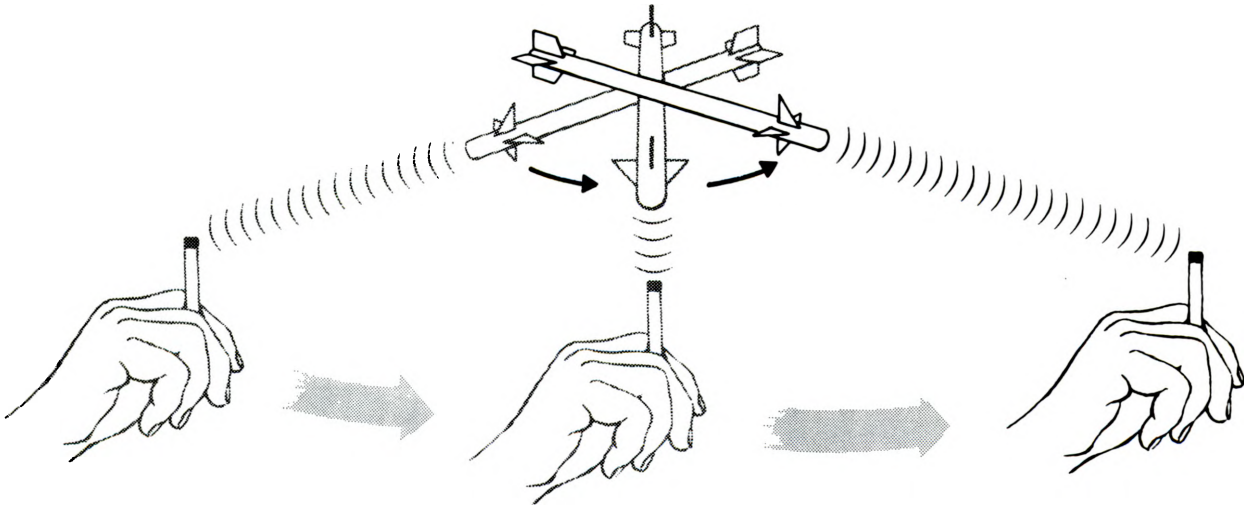
If any part of the IR detecting system becomes heated by absorbed energy, the part will reradiate the energy at wavelengths that differ from the original. If the detector is sensitive to these new wavelengths, the target will be obscured and ghost images will appear. Therefore an efficient cooling system for all components in the system is a necessity. In addition, the sensitivity of most detectors increases as ambient temperature decreases. The spectral response varies with changes in temperature, and the minimum detectable power necessary decreases with a decrease in temperature.

infrared seekers

general

Infrared seekers, sensitive to infrared (heat) radiation, perform the function of homing a missile to allow interception of heat-radiating targets. Infrared seekers operate passively because the target emits the homing

stimulus. Consequently, the missile does not reveal its own position, as it does with radar and other microwave beam-tracking devices. However, less information can be extracted from an infrared signal, thus the choice of seeker will be a military decision based on countermeasure techniques employed by the enemy.

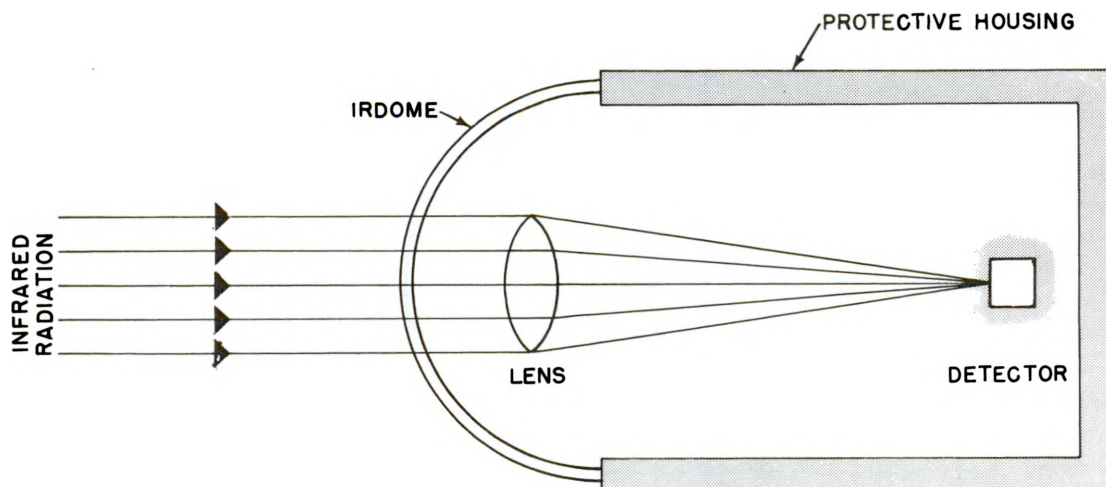


optical systems

Since light is an electromagnetic radiation, the emission of light from any source is a result of acceleration of electric charges. The sources of light with which we are principally concerned may be divided into two classes: 1) thermal sources, in which the radiation is the result of high temperature and 2) gaseous sources, which depend upon an electrical discharge through the gases. Optical devices are utilized to collect and focus this radiation on an infrared sensor. The heart of an optical system is the lens and the filter that is utilized with it.

A simplified processing system consists of a lens, a window or infrared radiation dome (irdome), composed of materials transparent to IR (such as optical glass, quartz, rock salt, germanium or silicon), and a processing unit used to protect the optical system from environmental forces that may be exerted on it.

The lens is an optical component consisting of one or more pieces of glass or material substances transparent to the radiation being used.



a simple IR optical system

CLASSIFICATION

Optical systems are classified basically as variable-focused or fixed-focused type. In variable-focus types, an image of constant size is attained, independent of the range. However, the mechanical problem of adjusting the position of the lense during missile flight is great. A more important deterrent for this type of operation is the loss of image resolution that occurs with a decrease in range. Once the range crosses the theoretical line of maximum resolution in a decreasing manner, resolution is lost and the fidelity of the reflection is decreased. In a fixed focus type of optical system, image size varies with range, while the flux density or energy or power passing through a surface remains the same. The disadvantage of this system is that a desired focal condition will be at an optimum value at only a pre-determined range.

LENSES

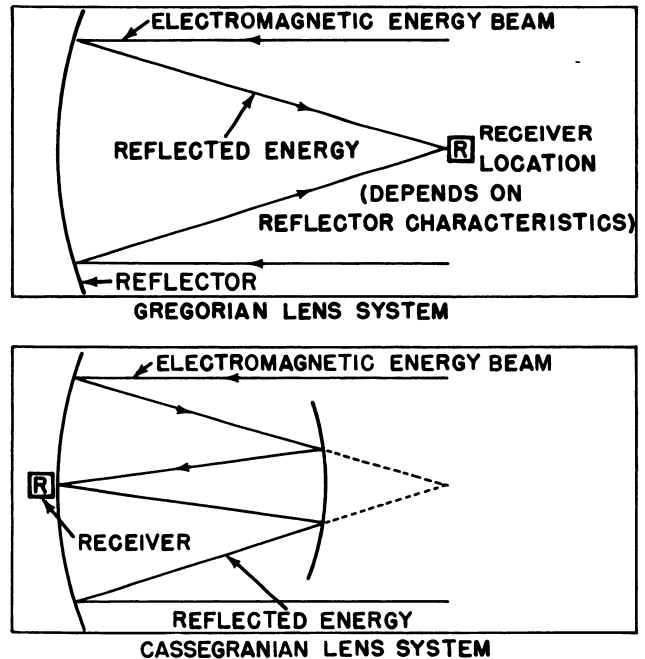
An IR lens is an optical component consisting of one or more materials transparent to the radiation being employed. Its surfaces are ordinarily spherical so that radiation received by the system will form an image of that object on an IR sensor material after passing through the lens. A lens used for IR transmission is designed transparent to a portion of the electromagnetic spectrum with a dielectric constant other than unity, and designed in a manner to produce a maximum indication of target presence.

frequency response characteristics

The frequency response characteristic of the material used in both the lens and the window is of extreme importance. The wavelengths that traverse the focus device must match the response frequency of the sensor. Lens and window materials are structually weak and often have a low resistance to thermal ranges and mechanical forces that they be subjected to. Often chemical changes will alter some of the basic properties of the materials, and special protective systems have been devised to protect the lens and window materials from both exterior and interior destructive forces.

cassegranian lens

A principal difference in lens systems is in the size of reflectors that must be used to provide the desired efficiency of a system. A cassegranian optical system employs a secondary convex mirror to reflect the light from the collecting mirror through a hole in its center. The first real image is found a small distance behind the collecting mirror. The result of the device is to decrease materially the size of the reflecting surface, allowing the packaging of a smaller, more efficient optical system.



A filter is a device for eliminating or reducing certain frequencies or bands of frequencies while leaving others relatively unchanged. Optical IR filters are employed to permit the passage of certain wavelength regions, such as atmosphere windows, enabling the target radiating characteristic to be more easily detected by the detector.

Infrared sources must be chosen carefully for operation in the NIR range. The shape of the eye sensitivity curve controls the sharpness of cut-off needed for NIR filters. A cut-off filter is used to filter the visible light from the source. Various densities of filters are required, depending on the source. Sources of NIR radiation, used either as target illumination or as modulated signal transmitters, always emit visible light along with the NIR waves. To make these sources invisible to the human eye, the visible frequencies must be reduced by filters, which unfortunately alternate the NIR frequencies. For any particular application, the choice of a filter is generally the minimum layer that reduces the visible light to a value reasonably safe from detection.

The maximum range of any detector depends in part on the intensity of the IR radiation released by the source. That characteristic of the filter that determines the passage of infrared waves is termed its effective hololuminous-transmission (ehT); it should be as high as possible. The characteristic of a filter that passes visible light is its effective visual transmission (evT), which should be as low as possible. In a practical situation, a filter that provides low evT probably has a fairly low ehT, so that the military requirements for security often limit the range that can be attained with any system. The frequency at which peak sensitivity of the receiver occurs also affects the type of filter chosen for use with a given source.

modulation

Electromagnetic radiation systems normally generate a carrier at a desired frequency, which is then modulated to carry the intelligence superimposed on it. Theoretically, the difference between a microwave antenna to transmit radar pulses or an infrared optical system are in degree only; the distinguishing feature is the wavelength of the electromagnetic radiation.

The purpose of a chopping reticle in an infrared search system is to provide or select the carrier frequency of the transmitted radiation. A chopper is usually a rotating disc with transparent and opaque areas, which cut off or permit radiation impulses at a rate dependent on the rotational characteristic of the reticle.

Scanning rate, caused by revolving mirrors which shift the focal point about the sensor to determine target location, modulates the energy beam at a rate equal to the frequency of the mirror's rotation. Hence the modulation frequency is the number of revolutions of the beam per unit time. When the scan results in a periodic interruption of a steady beam, the modulation frequency depends on the number of sectors in the mirror system. A high modulating frequency allows discrimination between target and background radiation and permits ease of separation. Because background radiation is normally of a lower frequency, a system that utilizes a high frequency of operation can detect the presence of the target more readily. In this respect, a slotted disc type scanner has an advantage over a mirror scanner.

infrared receivers

An infrared receiver can be fairly simple in design. Therefore it is possible to obtain a high degree of reliability in its operation. The essential components of an infrared receiver are a detector or sensor, a phase discriminator, a system of reference for direction guidance, a computer, and an output circuit capable of activating control devices.

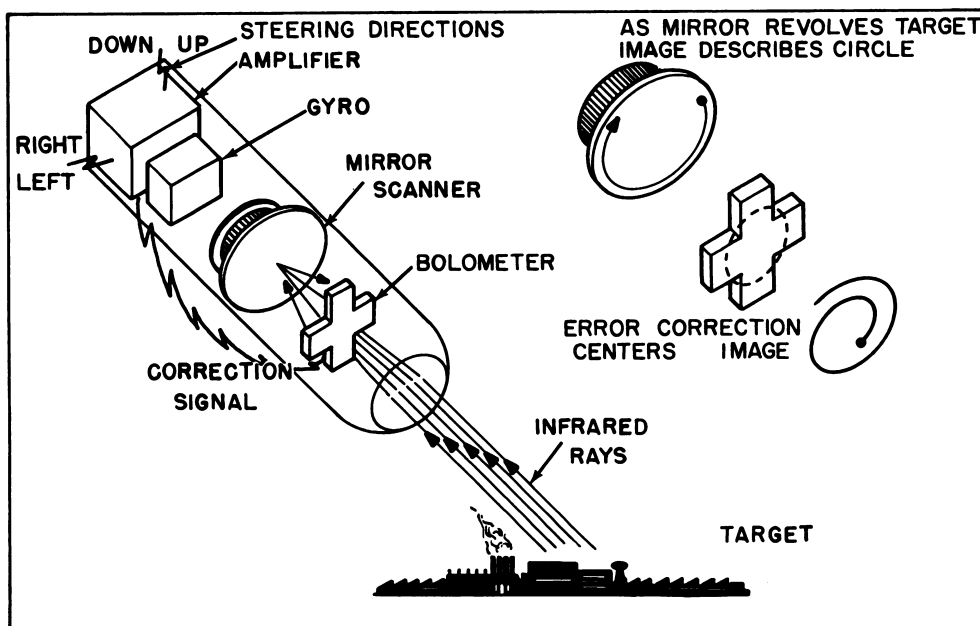
DETECTORS

The detector consists of an IR sensor, an optical system and a scanning system.

IR SENSOR

The target seeker, as illustrated, consists of a gyroscope and a folded reflecting telescope which spin together in the optical axis of the telescope. This gyroscope performs four essential functions.

- a. It provides a stable platform for the seeker telescope.
- b. It rotates the image chopper which transforms the infrared radiation received from the target into a pulsating signal.
- c. It provides the necessary torque to keep the axis of the telescope pointed in a fixed direction in inertial space.
- d. It measures the angular rate of the LOS.



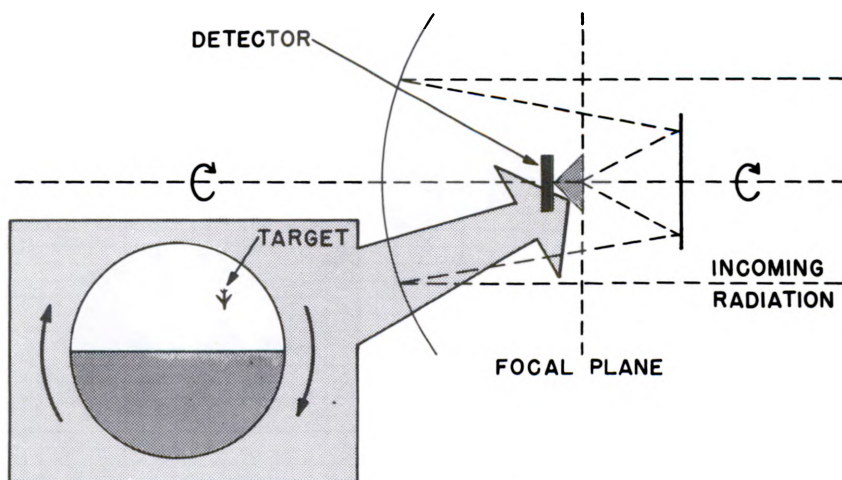
GUIDANCE SYSTEM USING A BOLOMETER

The purpose of the seeker telescope is to form an image of the target and to remain pointed at the target at all times. The folded reflecting mirror of the optical system focuses the received infrared radiation from the target upon a rotating reticle. The reticle is alternately opaque and transparent. The light transmitted by the reticle falls on a lead sulfide cell. The lead sulfide cell has an electrical resistance which is proportional

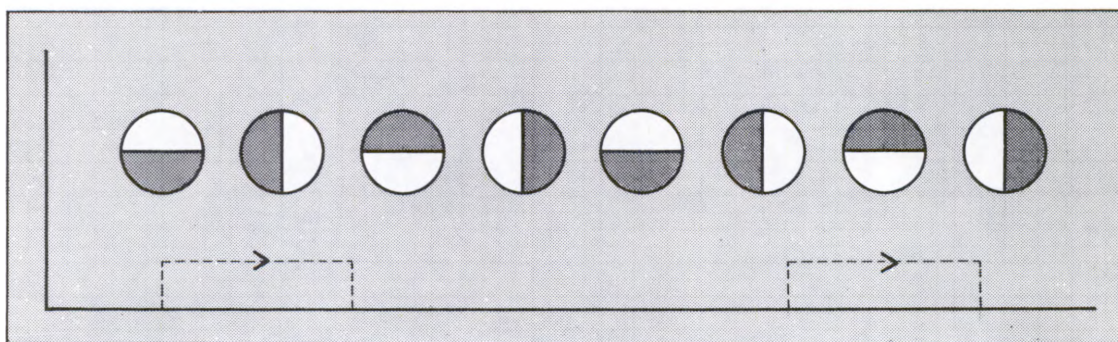
to the amount of received infrared energy.

To provide electrical signals that can be processed by the missile's circuitry, the optical system forms the radiation into an image and chops it, producing a pulsed radiation signal. The chopping action is accomplished by the reticle, as illustrated. The current from the lead sulfide cell is related to the rotational position of the reticle.

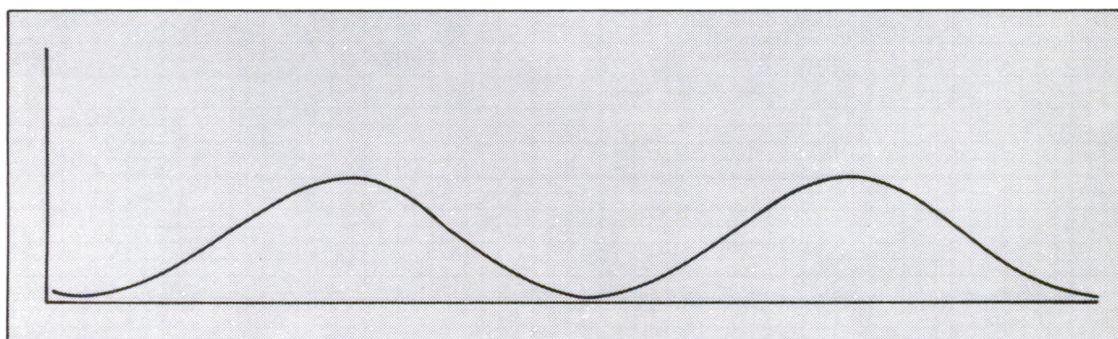
Whenever the target image is wholly on the clear portion of the reticle, the resistance of the lead sulfide cell is a minimum (maximum current). The angular position of the target relative to the telescope axis determines the time of maximum current flow through the cell.



The circuit rounds off the resulting current pulse and the signal appears as shown.

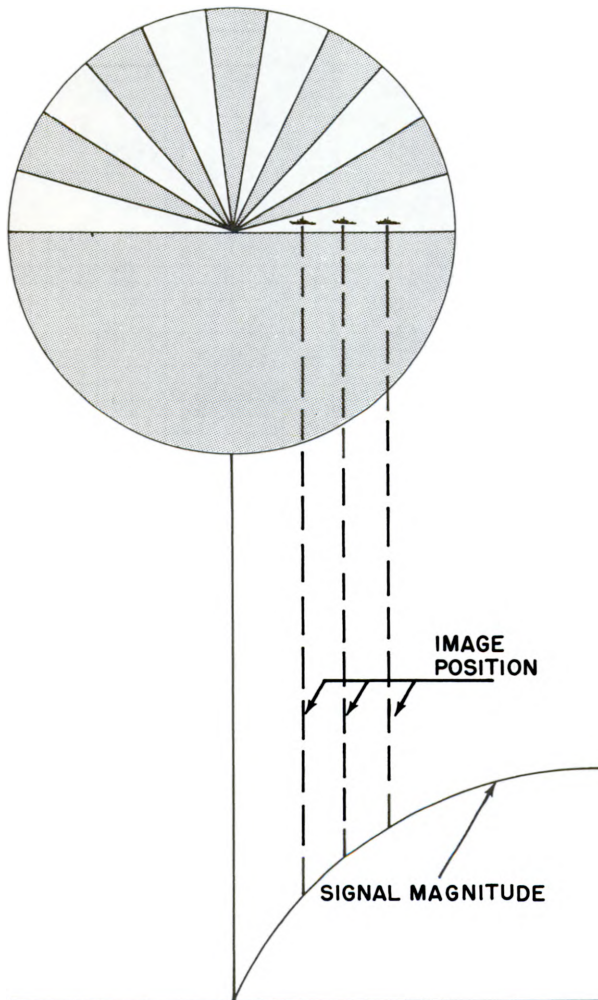


DETECTOR CURRENT



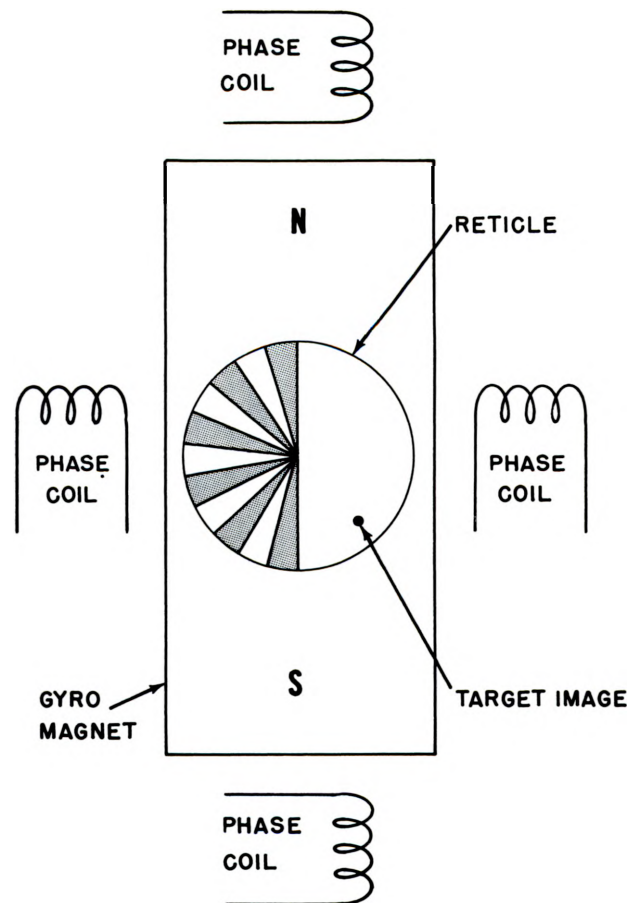
CIRCUIT CURRENT

The error signal discussed so far gives an indication of the direction the LOS is moving, but not its rate of motion. The seeker head, therefore, must be able to develop the LOS rate of change information which may be done by varying the amplitude of the error signal. This is accomplished by covering the clear half of the reticle with alternating clear and opaque sectors, as illustrated.



The signal strength developed in the lead sulfide detector cell is now dependent on how far the target image is from the center. The closer to the center, the less infrared energy is passed by the reticle. The further from the center, the more infrared energy is passed. The signal from the lead sulfide detector cell is amplified and, still containing both phase and amplitude information, is fed back to the solenoid coil around the seeker gyro. This produces an alternating field along the axis of the missile with the same frequency at which the gyro is spinning. This alternating field acts on the poles of the spinning magnet and results in a torque on the gyroscope proportional to the current in the coils. The resultant torque processes the seeker lead toward the target until the seeker axis and LOS are aligned. Thus, the seeker lead is able to track the target continuously.

The error signal that processes the gyro is sent also to the amplifier of the control system as a command signal. Before it can be used by the control system, however, it must be converted from its original space stabilized polar coordinate system to the missile reference system. This is accomplished by the phase discriminator. Four phase coils are spaced at 90° intervals around the seeker gyro case and attached to the missile airframe. These coils cause the permanent magnet, an integral part of the gyro, to register its angular position relative to the missile body. The angular position is in terms of currents generated in these coils. The phase discriminator compares the currents from the phase coils with the error signals. This comparison results in a conversion of the error signal into orders for the control system in the missile reference system. The control system is then amplified by the navigation ratio and the modified signal activates the servos to produce the required missile turning rate. When the missile has the proper lead, there will be no rotation of the LOS, and thus, no turn signal to the missile control system. The missile will then be on a collision course.



design parameters

The design of an IR detection system must take into account many characteristics previously discussed. Of primary importance is the target's frequency (wavelength) of transmission. It is also necessary to know maximum operating range desired of the equipment. Background and ambient radiating levels will also affect the design characteristic. Noise levels must be determined so that necessary power requirements can be met. This characteristic will affect the selection of window, detectors, optical system, carrier frequencies, and modulation systems. Infrared detectors differ significantly with respect to spectral response, sensitivity, response time, etc. Thus, the most important factor in design is to determine the target's characteristics in the medium it is traversing. Only then can the receiver be designed for optimum reception results.

military applications and considerations

There are a number of advantages in using infrared radiation for target detection over other forms of electromagnetic transmission:

- 1) IR systems are passive
- 2) Military targets are excellent sources of IR
- 3) Complete jamming of IR systems is very difficult
- 4) IR systems can be engineered smaller and less expensive than comparable systems
- 5) IR systems have higher resolutions

The above advantages make IR systems a good choice for many military applications, including: photography and aerial mapping, missile guidance, fire control, communications, and flying.

Disadvantages of IR are the slowness of the reaction time, and the low operative limit of effective operation. Environmental factors (fog, rain, etc.) limit effective operations and decoy heat sources near targets is an effective countermeasure and may cause false tracking. Considering the altitudes at which some detector may have to operate, the heat sealing device may not be sensitive enough to discriminate between target and background.

From the standpoint of countermeasures, it is not easy to detect the operation of a passive system (impared) compared to the ease of location of an active system (radar). Countermeasure technology is much further advanced when dealing with microwave transmission than with IR. Heat or high energy flares of proper wavelength must be ignited and displayed at just the right time to be effective in distracting the attacking missile.



R A D A R

Radar is a term derived from RAdio Detection And Ranging, and signifies a method whereby radio waves are employed to detect the presence and location of objects in space. In the early 1900's it was discovered that electromagnetic radiation incident on any discontinuity (radiating surface) in a medium traversed, is reflected to a greater or lesser extent, depending on the radiation coefficient of that surface, and the power contained in the transmitted beams.

Target detection is accomplished by transmitting a beam of high-frequency energy over the area to be searched or scanned. When the energy strikes a reflecting surface, a small portion of the energy is reflected or reradiated toward the original radiating source, where a sensitive receiver located near the transmitter detects the echo pulse and therefore the presence of the target.

To determine an object's position in space completely, it is necessary to find its three positional coordinates (range, azimuth angle, elevation angle) within a required frame of reference.

radar spectrum

Although the principles of radar detection have been applied over the whole radio spectrum (from 10 kc to 10^9 mc), the high-frequency or microwave end of the region has been used primarily for military radar equipment development. Radar generally is considered to cover the frequency range of approximately 0.390 gc (gc = gigacycles = 10^9 cycles), or wave lengths of from 76.9 cm to 0.536 cm. The long wavelength region is used for ground radar; the center region is used for airborne radar; and the shortwave region is used for guided missile application. This short microwave region borders on the far infrared region, which is the reason why components of radar and IR systems (particularly antenna and lens systems) are similar.

applications of radar

An important factor in the military application of radar is that the units of a radar set are completely self-contained and provide an almost instantaneous flow of information concerning objects in their search area. In a few seconds, a radar set is able to produce a multiple display of enemy aircraft, surfaced submarines, and fixed objects located hundreds of miles away. The complete survey can be carried out so rapidly that a plane traveling at the speed of sound and located 100 miles from the radar set would travel only three feet from the instant of initial detection to the instant at which the next interval of target surveillance began. Thus, radar can determine true target position at any instant of time, regardless of target speed.

A weapon system must not only locate an object, but identify it, determine its flight path, and predict its final objective with a high degree of reliability. Detection, identification, tracking, computation, and interception must take place before the release of enemy destructive energy can be accomplished. Long-range radar detection devices are generally employed in early warning systems, usually located hundreds or even thousands of miles from priority target areas. The weapon system is thus given time to determine enemy missile flight paths and to set in motion interception and destruction devices. To insure interception and to increase kill probability, more sensitive radar is then utilized at shorter ranges, permitting more accurate determination of target coordinates. Often this radar is airborne and is used to guide interceptor craft to the target areas. Whether the sensitive-type radar is ground controlled or airborne, the integrated use of all types of radar systems is an important aspect of a country's defense posture.

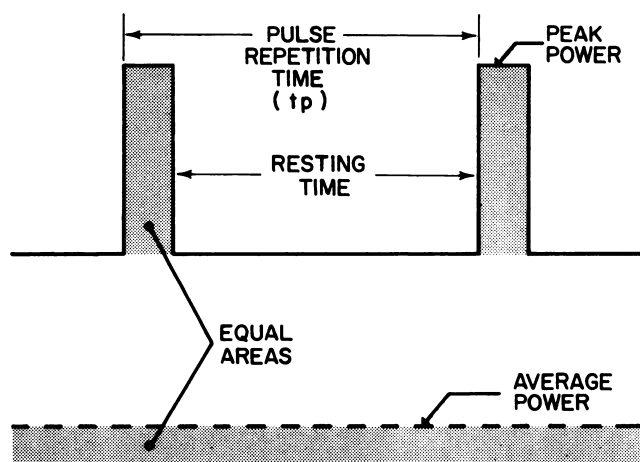
TYPES OF

Radar systems employ two basic types of energy transmission: pulse and continuous wave (CW). The basic principle of pulsed radar is that the transmitter sends out radio waves in a series of short, powerful pulses and then rests during the remainder of its cycle. During the period in which the transmitter is at rest, echo signals may be received and timed to determine the range to the reflector surface. In CW radar, on the other hand, the transmitter sends out a more or less continuous signal. If a nonmoving surface is in the path of the transmitted wave train, the frequency of the reflected signal will be the same as that of the transmitted signal. If the surface is moving, the frequency of the reflected echo will differ from that of the transmitted signal and the frequency difference can be used as an indicator of target motion. In CW transmission, either a movement of the radar or the target is necessary to produce an indication of target presence.

ENERGY TRANSMISSION

pulse transmission

Pulse radar makes it possible to measure range in terms of the time intervals between transmission of a pulse and the reception of an echo or reflected pulse from the target. The relationship of these pulses is shown.



The pulse repetition time (t_p) must be of sufficient duration to allow the echo pulse to return from the maximum range of the system, otherwise the reception of an echo will be obscured by the succeeding pulse. The total time required for the completion of a cycle (the pulse duration plus the resting time) is called the repetition period. The reciprocal of repetition time is the repetition frequency (usually expressed in gigacycles). The necessary time delay determines the maximum frequency which can be used for pulse repetition. The minimum range at which a target can be detected is determined by the pulse width, T_d . If a target is so close to the transmitter that the echo is received before the transmitter is cut off, the echo reception will be almost completely attenuated by the width or time duration of the transmitted pulse. Because the pulse duration time must be short to increase reception of nearby targets and yet contain sufficient power to insure a return echo of sufficient magnitude from the maximum range of the set, extremely large transmitted power outputs are required to produce a pulse of sufficient energy. The useful power of the transmitter is contained in the radiated pulses and is termed the peak power of the system.

Since the radar system is not transmitting for a long period of its total cycle, the average power is quite low, when compared with the peak power during the pulse time. The relationship between average power dissipated over the entire cycle and peak power developed during the pulse time can be expressed by the following equation:

$$\frac{\text{average power}}{\text{peak power}} = \frac{\text{pulse width}}{\text{pulse repetition time}}$$

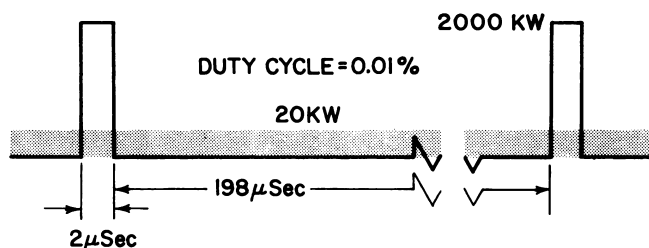
The greater the pulse width, the higher the average power; the longer the pulse repetition time, the lower the average power. The relationship between pulse width and pulse repetition time and between average power and peak power may be expressed in duty cycles and represented as:

$$\frac{\text{pulse width}}{\text{pulse repetition time}} = \frac{\text{average power}}{\text{peak power}} = \frac{\text{duty cycle or}}{\text{duty ratio}}$$

For example, a 2-microsecond pulse repeated at a frequency of 5000 cps, represents a duty cycle of .01, since the time for one cycle is 1/5000 of a second, or 200 microseconds.

$$\frac{2}{200} = .01 = \text{duty cycle}$$

Assuming a peak power of 200 kilowatts, then for 2 microseconds, 2000 kilowatts are available. Since average power = peak power x duty cycle then the average power = 2000 x .01 kilowatts.



range determination

The effectiveness of range determination depends primarily on the ability of the system to measure distance in terms of time. Electromagnetic radiation travels in space at a constant velocity of 186,300 miles per second. When it is reflected there is no loss in velocity, but merely a redirecting of the energy path. Range is then determined by the time required for a two-way energy transmission. The determination of range can be made by using the equation

$$R = \frac{c}{2t}$$

where R = the distance or range from radar set to target
 c = the speed of propagation of radio waves in air
 t = the time required for the two-way trip

The velocity of electromagnetic radiation is approximately

- 1) 186,300 miles/sec
- 2) 984 ft/ μ sec
- 3) 328 yds/ μ sec
- 4) 6.18 nautical miles/ μ sec (using the approximation of 6000 ft equals one nautical mile)

Since range determinations are based on the measurement of time required for a pulse to travel to a target and return, radio waves traveling 186,300 miles in a second, for example, will travel 0.1863 miles in a microsecond. This corresponds to a time of 5.375 microseconds to travel one mile. If an aircraft or other reflecting object exists at a distance of one mile

continuous wave transmission

When radio-frequency energy which is transmitted from a fixed point continuously strikes an object which is moving toward or away from the source of the energy, the frequency of the reflected energy is changed. This change in frequency is known as the Doppler effect. The difference in frequency between the transmitted and the reflected energy determines the presence and the speed of the moving target.

When the source and target are both stationary the frequency, f , wavelength, λ , and velocity, c , of a wave are related by the expression $c = f\lambda$. If the source is moving with velocity s , as illustrated, the distance between crests (λ') is increased from c/f to $\frac{c+s}{f}$ in the direction to the left of the source. Similarly, the wavelength is decreased to $\frac{c-s}{f}$ in the direction to the right of the source. To an observer standing to the right of the source, the frequency heard is $\frac{c}{\lambda}$, which may be rewritten as $c \frac{f}{c-s}$ or $\frac{c}{c-s} f$. If the observer moves toward a stationary source with velocity s , the frequency heard is raised from $\frac{c}{\lambda}$ to $\frac{c+s}{\lambda}$. Since the original wavelength is expressed by $\lambda = \frac{c}{f}$, the new frequency may be rewritten as $(c+s) \frac{f}{c}$ or $\frac{c+s}{c} f$. If a radar transmitter-receiver is moving toward a stationary target, the frequency of the reflected signal is raised because of the motion of the transmitter and the receiver. The transmitted frequency is raised from f to f' by the motion of the transmitter, and from f' to f'' by the motion of the receiver, as follows:

$$f' = \frac{c+s}{c} f \text{ and}$$

$$f'' = \frac{c}{c-s} f'$$

Therefore the resultant frequency is:

$$f'' = \left(\frac{c+s}{c} \right) \left(\frac{c}{c-s} \right) f$$

$$= \frac{c+s}{c-s} f$$

The fraction $\frac{c+s}{c-s}$ may be reduced to simpler terms:

$$\frac{c+s}{c-s} = 1 + \frac{2s}{c-s}$$

Since the speed of the fastest missile is much less than c (the velocity of light), only a negligible error is introduced by assuming that $2s/(c-s) = 2s/c$.

Therefore $f'' = (1 + \frac{2s}{c}) f$

This equation demonstrates that the received frequency changes linearly with radial velocity. It does not matter whether the radar, the target, or both are moving as long as s is the relative velocity along a line joining them.

For a moving target, the Doppler shift frequency, Δf , is given by

$$\Delta f = \frac{2fs}{c}$$

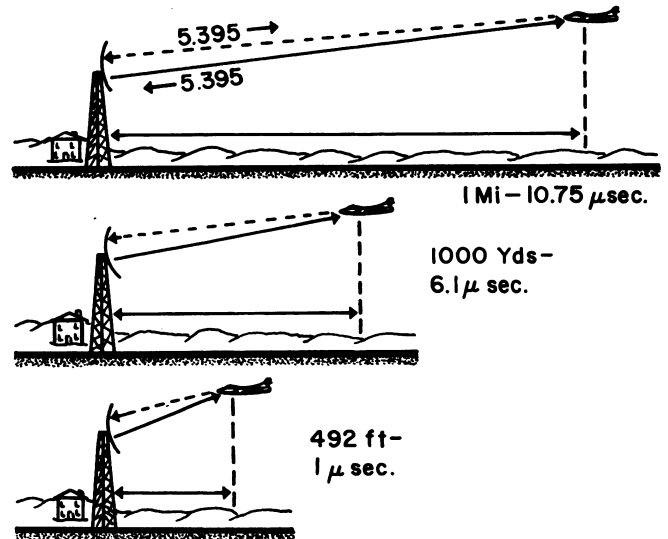
since $f = \frac{c}{\lambda}$ then $\Delta f = \frac{2s}{\lambda}$

The Doppler shift has important operational applications in shipboard to determine target true motion or aspect. If the ship carrying the radar equipment is in motion, the echoes received from a stationary target differ in frequency from the transmitted signal according to the above equation, except when the signal is transmitted on a beam where there is no component of the ship's motion along the path of propagation. The Doppler shift is greatest for targets dead ahead or dead astern and decreases as the direction of transmission is shifted toward the beam. If both the target and the echo-ranging equipment have motion, the velocity used is that component of the target's velocity relative to the echo-ranging equipment which lies on a straight line connecting the two. If this component is zero, i.e., the target is moving parallel with its seeker, then there is no Doppler effect.

Since the Doppler frequency is a measure of velocity only, there is no way of measuring range directly at a

from the radar set, the reflected echo will reach the receiver 10.75 microseconds after transmission. This interval is the round trip time and can be converted to linear range measurements.

More sensitive radar sets determine distance in yards instead of miles. Since the constant velocity of electromagnetic radiation is 328 yards per microsecond, a target located 32,800 yards from the source will require a round trip time of 200 microseconds. Since there are 1760 yards in a mile, the round trip time for a distance of one yard is 0.0061 microseconds. A more convenient figure is the round trip echo time for an object 1000 yards away, 6.1 microseconds.



constant carrier frequency. Tracking with CW radars is possible theoretically to zero range.

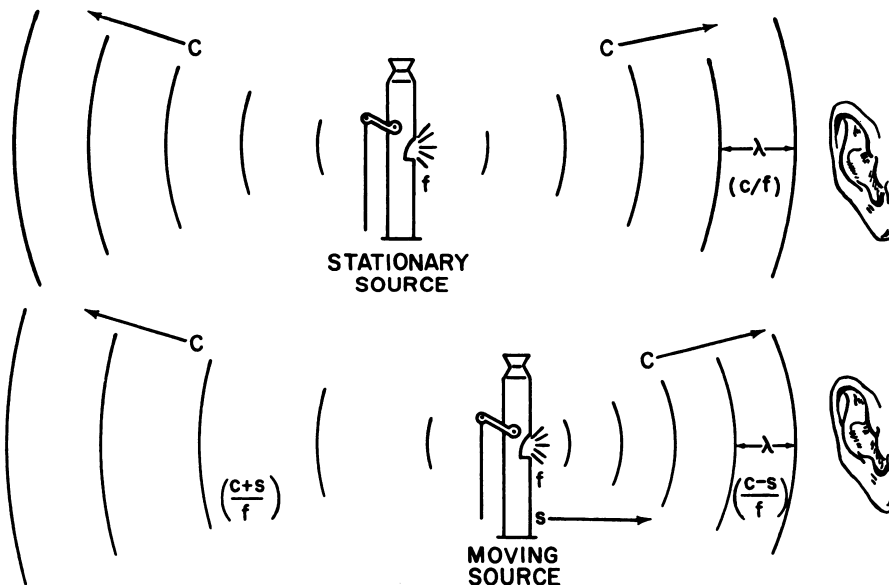
If a target is moving, its radial velocity, relative to the radar can be detected by comparing the transmitter frequency with the echo frequency, which will differ because of the Doppler shift. The difference or beat frequency, f_d , sometimes called the Doppler frequency, is related to target velocity, v , and transmitted wavelength λ (in cm) as follows:

$$f_d = 89.4 v / \lambda$$

To track a target with CW Doppler radar, range information is required, but f_d is not a function of range. However, if two separate transmitters operating at two frequencies f_1 and f_2 are employed, then the relative phase

between the two Doppler frequencies is a linear function of range to the target. In this system a hybrid mixer is used to combine the two transmitted frequencies and separate the two received frequencies to permit the use of one transmitting and one receiving antennas.

Instead of two transmitter frequencies the same result is achieved by sweeping the transmitter frequency uniformly in time to cover the frequency range from f_1 to f_2 . The beat frequency f_b between the transmitted and received signals is then a function of range. In this type of radar the velocity as well as range is measured so that the velocity information, which is rate of change of range, can be used to control range and velocity gates for range tracking. Conical scanning is also used for angular tracking.



pulse-doppler radar

The outstanding advantage of pulse radar is that the rate of collection of information is very high. The rate at which information is collected is a function of the rate at which separate echo pulses can be received. Theoretically, this can be as high as $\frac{1}{t}$, where t = pulse duration. The essential advantage of the high information collection rate can be realized by the use of narrow search beams and high peak power. However, as pulse duration time is decreased, IF bandwidth requirements in the receiver are increased, and system noise becomes a serious obstacle. Another limitation is that narrow beam pulse radar is extremely sensitive and picks up unwanted reflections from sea clutter, storm clouds, and other miscellaneous objects. When the strength of these echoes is as severe in intensity as the strength of the desired signal the indication is unable to distinguish between the wanted and unwanted signals.

When this occurs, the only method that can be used to distinguish between a target echo and an equally strong echo from the background is to utilize the Doppler principle as a determinant of target motion with respect to its surroundings.

Doppler pulse radar systems, as well as CW systems, can use Doppler to distinguish between stationary and moving targets. The arbitrary distinction between CW and pulse radar systems may be used to distinguish between a CW pulse-modulated Doppler system and a pulse radar moving target indicator (MTI) system. A radar that transmits more than 10 percent of the time is considered a CW radar. The systems are fundamentally very similar, and the distinction between them is made only for the purpose of classification.

Pulse radars may be modified in one of several ways to employ the target Doppler frequency to detect a moving target. The Doppler shift effects prt, pulse width, and carrier frequency, and a pulse radar can be designed to recognize one of these three effects. The presence of a moving target in a background of external noise is known from the amplitude variations it produces in the noise spectrum. Amplitude detection employs non-coherent reception. Moving targets are also detected by the phase differences between the target signal and noise components. Phase detection employs coherent reception. In noncoherent detection, the local oscillator of the receiver does not have to be frequency stabilized, since phase changes are not involved. If the external noise level present is not too high, the target may be distinguished visually from the noise clutter on an A-type indicator. The main advantage of non-coherent detection is that it is insensitive to relative motion. Thus, the receiver can be carried by a missile or aircraft without the complications of phase shifts.

In coherent detection, a stable CW reference oscillator produces beats with the target signals either at radio frequency or intermediate frequency. The reference oscillator is locked in phase with the transmitter during each transmitted pulse by means of a very stable local oscillator which feeds the mixers. Mixer 1 produces the conventional 30- or 60-mc IF to the receiver, while mixer 2 reduces the RF locking pulse to intermediate frequency. By comparing the signal echoes from the output of mixer 1 with the CW signals from the reference oscillator, beats result which depend only on the phase difference between the two oscillators. Since the reference oscillator and transmitter are locked in phase, the echoes, in effect, are compared with the transmitter.

The phase relationship of the echoes from fixed targets to the transmitter is constant; therefore the amplitude of the beats remains constant. Beats of varying intensity indicate a moving target, because the phase difference between the oscillators changes as target range changes.

The video output of the receiver, amplitude modulates a magnetron oscillator and amplifier of relatively low frequency (10 to 20 mc) and is divided then by a circuit parallel to the cancellation circuit. The time delay of the upper line of the parallel circuit is matched exactly to the prt. Since the signals from fixed targets do not cause phase changes, they cancel out, because the polarities of the detectors in the delayed and undelayed lines are opposite in sign. The signals from a moving target, which cause phase difference, do not cancel, and are transmitted to a PPI for visual display.

frequency-modulated radar

Frequency-modulated (FM) radar systems transmit a continuous frequency-modulated signal. A target reflects a delayed signal which at any one instant differs in frequency from that of the transmitter by the amount of frequency change caused by the target's position and motion. The difference frequency depends on the delay interval, and therefore is a function of range.

Separate antennas are employed for transmission and reception so that no time delay is introduced for switching a dual-purpose antenna. This permits measurements of ranges as short as a few feet. The bandwidth, the rate of information collection, and the resolution are much smaller in CW transmissions than for pulse radar. Therefore, radar is useful primarily where one important target is illuminated and where speed of indication is not critical. The principle application of radar in the Navy is in the radio altimeter, where reflection occurs from the surface of the ground or ocean.

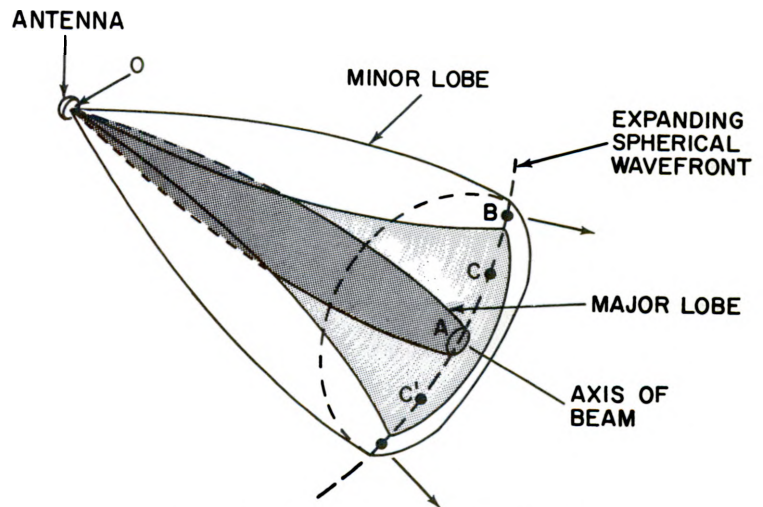
PRINCIPLES OF ENERGY TRANSMISSION

As a wave is propagated outward from the antenna, wavefronts form spherical surfaces that expand radially. If the antenna is directional (which is almost always the case) the distribution of power over the wavefront surface is not uniform. The power density (power per unit area) is greater at points within the beam (A) and small at points on the extremities of the beam (B). Points of no radiation are called null points (C). The power density decreases gradually from a maximum at point A to zero at point C. The region of main radiation between C and C' is called the major lobe, and regions of low radiation are called minor lobes. The major lobe is normally specified by beam width and is defined as the beam angle between half-power points. This angle may be specified as the angle between lines on opposite sides of the OA axis along which the power density is half as great as it is on the axis. The beam angle serves as a measure of the resolution of a set. The power distribution of a radar beam is based on the principle that power density is proportional to the square of the field intensity. This is expressed by the following equation:

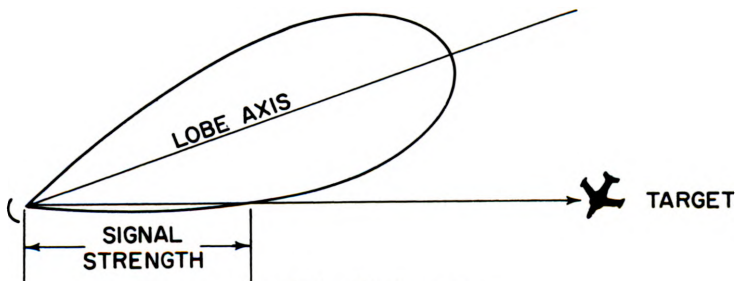
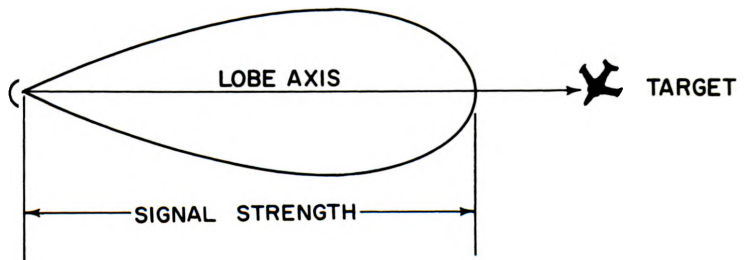
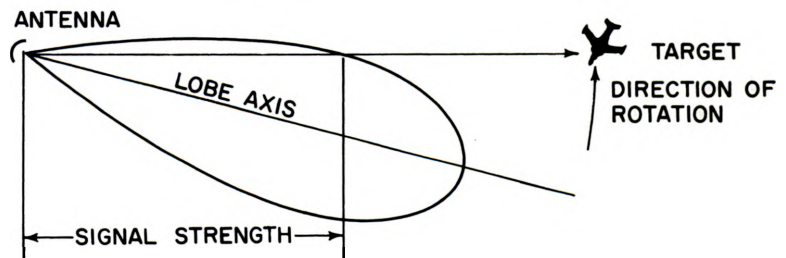
$$P = K \frac{\sin \left(\pi L \frac{\theta}{\lambda} \right)^2}{\pi L \frac{\theta}{\lambda}}$$

where θ = angle from beam axis
 L = antenna length
 λ = wavelength
 K = constant

To obtain the greatest accuracy in determining azimuth angles, the beam must be as narrow as possible. Because of diffraction effects, a narrow beam can be produced only if the dimensions of the radiating surface are large in relation to the radar wave. In the illustration shown, relative signal strength is plotted against angular position of the antenna with respect to the target. The maximum signal is received when the axis of the lobe passes through the target. The sensitivity depends on the angular width of the lobe. Use of truncated reflectors develops a lobe when the horizontal diameter of the beam is greater than the vertical. This is an excellent beam pattern for ship search radar. The small horizontal beam angle provides excellent azimuth resolution, and the large vertical beam angle negates the effects of ship roll and pitch.



radiation pattern from directional antenna

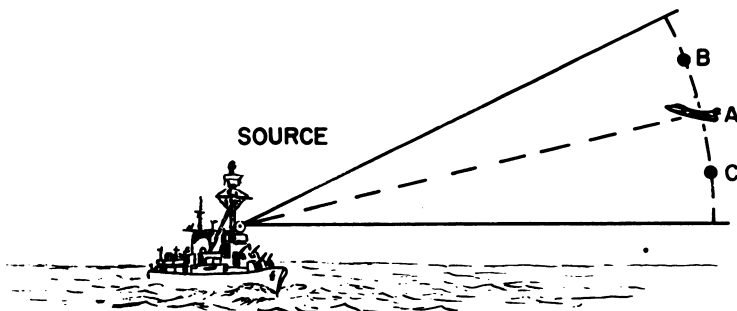


*relationship between
beam axis and target bearing*

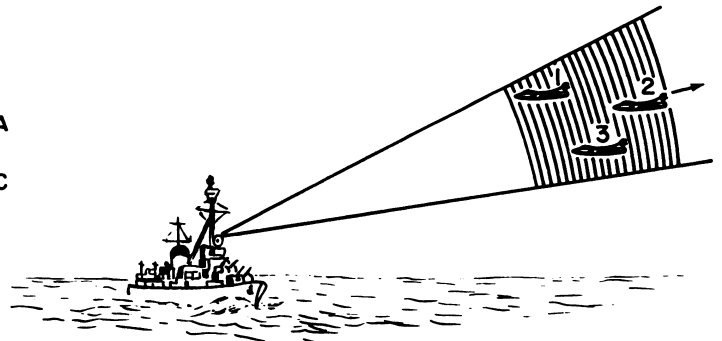
beam requirements

The tactical application of the radar determines the requirements of the beam. The types of beam required for target search is different from that required for tracking or for in-close interception applications. Because the purpose of an early warning radar is to search for and detect a target at maximum ranges, high transmitter power, wide beam patterns, and scan rates that are tied in with the area to be put under surveillance are the primary requisites. Accuracy of

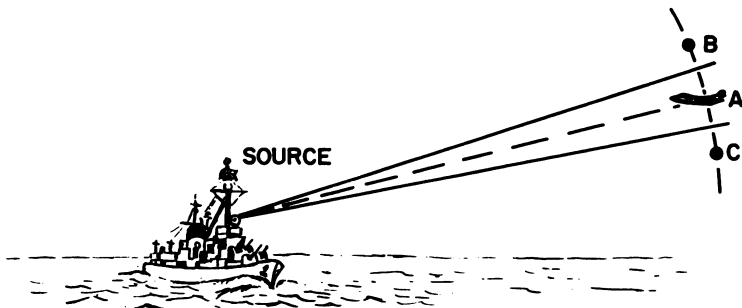
information output is sacrificed for maximum search probability. When used for tracking or guidance, operations that require extreme accuracy, the radar beam must be of narrow width propagated at higher frequencies and rotated in a chosen segment of the entire search area. The narrowness of the antenna beam determines the accuracy with which the radar can measure azimuth and elevation. It also determines the angular resolution of the radar, in the same manner as pulse duration determines range resolution.



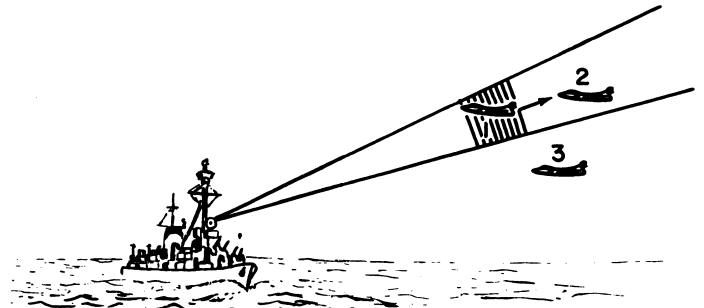
The airplane at A can move to B or C without producing a noticeable change on the radar viewing-screen.



A wide beam and long pulse produce an incoherent echo instead of separate echoes.



If the airplane at A moves to B or C, it can no longer be detected, unless the beam is moved.



Individual objects adjacent to one another can be detected if the radar set has a narrow beam and sends out a short pulse.

beam power measurements

The relation of maximum range to beam power characteristics is an important consideration. The peak power transmitted from an antenna (P_t) is divided over the spherical wavefront radiating outward from the antenna. The area of its spherical surface is $4\pi r^2$, and if the radiation were uniform, power density would be $P_t/4\pi r^2$. (The power density of a wave radiated from a point source is inversely proportional to the square of the distance from the source because the directional power distribution of the antenna is not uniform.) Power density is greater in the search area, and the power density formula must be implemented to account for the increase. Power density of the transmitted wave equals $(P_t/4\pi r^2) G_t$, where G_t is the factor of concentration of beam energy and is dependent on antenna array formation. G_t may also be expressed as the power gain of the transmitting antenna. It is often expressed relative to an isotropic antenna with a gain of unity.

When the transmitted wave strikes a target, it is in part reflected and, if the ratio of power reflected by the target to power impinging on the target is n , and θ represents target area, then:

$P_t/4\pi r^2 \cdot G_t \cdot \theta \cdot n$ = power reflected by target. Thus, the power density at the radar receiving antenna would be:

$$\frac{P_t}{4\pi r^2} \cdot G_t \cdot \theta \cdot n \cdot \frac{1}{4\pi r^2}$$

If A = effective area of the radar antenna, then:

$$P_r = \frac{P_t G_t \theta n A}{16\pi r^2} = \text{power delivered to the}$$

receiver.

Because the inverse-fourth-power law applies both to transmitted and reflected waves, echo strength falls off very rapidly with an increase in range. The maximum range can be determined if P_r represents the power of the weakest echo pulse that can be detected. When P_r is at a minimum, then $P_r = P_{r \text{ min}}/T$, where $P_{r \text{ min}}$ is the minimum energy level and T is the pulse duration. Where P_r is replaced by $P_{r \text{ min}}/T$, r becomes the maximum range.

$$r_{\text{max}} = \frac{1}{2\sqrt{\pi}} \sqrt[4]{\frac{P_t G_t \theta n A}{P_{r \text{ min}}}}$$

It is evident that large changes in P_t (transmitted power) will not greatly increase the range of the radar system. If, for example, the peak power is doubled, the range is increased only 19 percent ($\sqrt[4]{2} = 1.19$). The average power output, however, has sizeable effect upon maximum range. The pulse repetition rate effects the $P_{r \text{ min}}$ time as the rate of pulse appearance on the indicator screen increases the sensitivity of the system. The average power which is proportional to repetition frequency is vital in the determination of power requirements. The beam of energy transmitted by a radar always diverges or spreads to some degree. Also, the greater the distance to the object, the smaller the amount of energy reflected. As an object moves farther and farther from the radar set, a limit is reached beyond which the reflections are too weak to be detected by the sensor device. To increase the maximum range of detection, the strength or power output of the pulses transmitted must be increased. This can be accomplished by increasing the power output of the transmitter and is equivalent to illuminating the target more brightly.

PRINCIPLES OF REFLECTION,

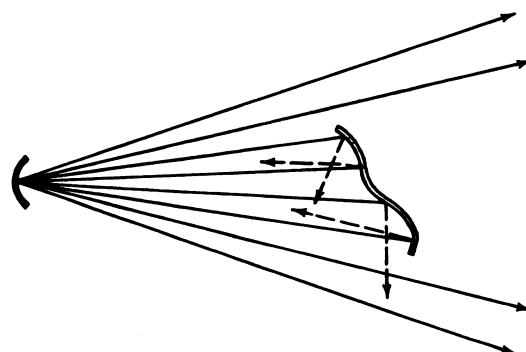
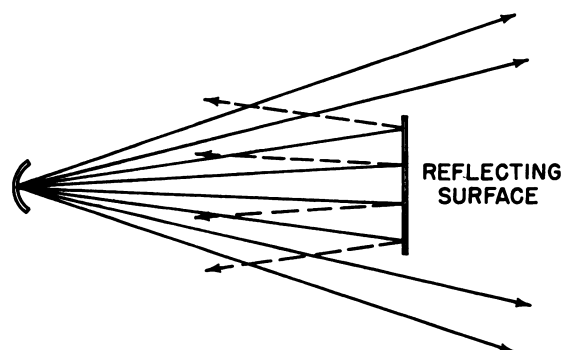
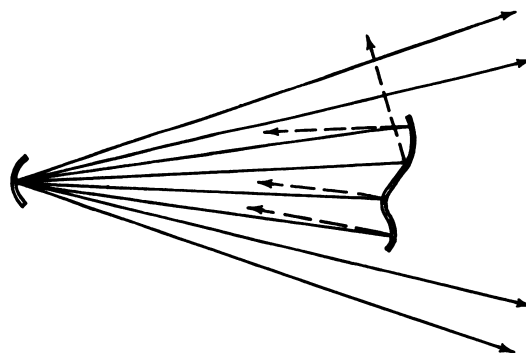
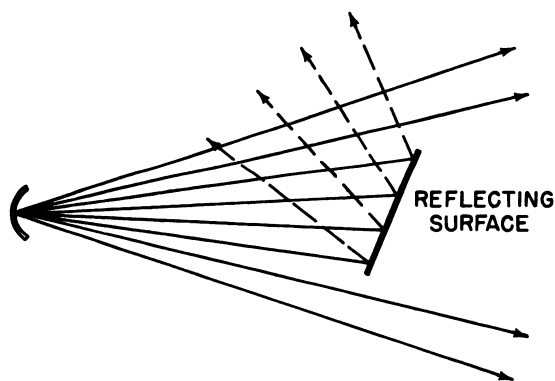
The radar transmitter produces the energy and the antenna radiates this energy into space, usually in the shape of a narrow beam. The antenna is then rotated so that the beam sweeps or searches the area required of it, and the radar gathers information about the distance and direction of movement of all depths within its average area. The reflection of the short wavelengths utilized by radar takes place from almost all objects in their paths.

Airplanes, missiles, ships, water, buildings, or practically any material substance will reflect these short waves. Fortunately all objects do not reflect these waves equally well. The intensity of the reflected echo depends upon the substance of which an object consists and also upon the size and shape of that object. From almost any shape or from any position of an irregular surface some of the reflected energy will be returned to the source, and can be used to identify the target position.

reflection

When a radiated electromagnetic wave encounters a conducting surface, reflection of energy from the surface occurs. Reflection from the surface takes place in accordance with the law of reflection which states that the reflected and incident wave trains travel in directions which make equal angles with the normal to the reflecting surface and are in the same plane with it. These angles are called the angle of reflection and the angle of incidence respectively. For normal, incidence, both angles are zero. Uneven surfaces reflect in a multitude of directions, and such reflection is said to be diffuse. Reflection can be expressed in terms

of the reflection coefficient of a surface, or the ratio of reflected field intensity to the incident field intensity. In radar, the most important consideration is the ratio of electric field strength of the reflected wave to the incident wave. Often incident energy is lost because of the presence of natural obstacles in the path of the radiation. Dust, snow, or water vapor will scatter the incident radiation in a haphazard manner, resulting in a loss in beam power. By far the greatest loss in field intensity which occurs is a result of diffusion caused by roughness and irregularities in the conducting surfaces.

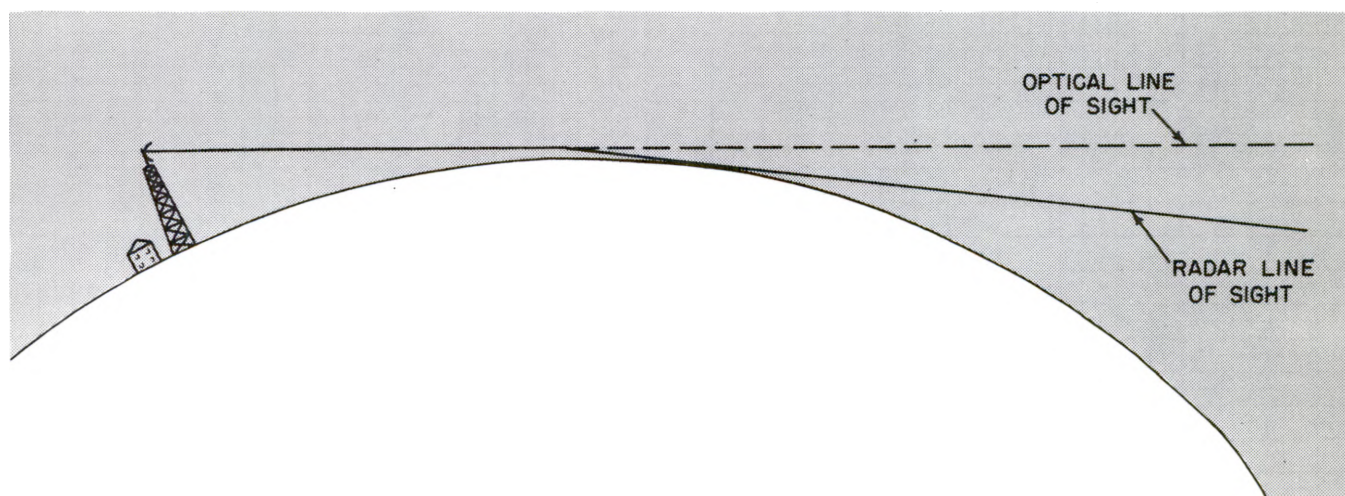


REFRACTION AND DIFFRACTION

refraction

Refraction is the bending of a radiated beam as it travels through space. The law of refraction, usually called Snell's law, states that the refracted wave lies in the plane of incidence, and the sine of the angle of refraction bears a constant ratio to the sine of the angle of incidence. If the medium being traversed is uniform, the velocity of propagation is constant and the

beam travels in a radial manner. Earth's atmosphere is not a uniform medium; the density of the air at lower altitudes results in an increase in the index of refraction. Thus, there is a downward bending of the rays of propagation, as shown. Because of the linear increase in the index of refraction with an increase in altitude, velocity of propagation increases also.



*extension of radar horizon
as the result of atmospheric refraction*

diffraction

Diffraction occurs when a wave passes through a restricted aperture and spreads out in a pattern determined by the ratio of the wavelength to the diameter of the aperture. For a single opening of width a and wavelength λ falling on the opening at normal incidence, the intensity of the wave at angle θ can be expressed by

$$I = \frac{R_0^2 \sin^2 \left(\frac{\pi a \sin \theta}{\lambda} \right)}{\left(\frac{\pi a \sin \theta}{\lambda} \right)^2}$$

The importance of this principle in lens or antenna reflecting systems will be more fully explained later in this chapter.

Since both earth and water reflect short radio waves, targets located underwater or below the surface of the

ground cannot be detected by radar. Since ultra-high-frequency waves travel through space along nearly straight lines, radar cannot see around the curvature of the earth; objects below the horizon cannot be detected. Actually the horizon distance for radar waves is somewhat greater than the horizon distances for light waves because radar waves bend slightly in their motion through the atmosphere. The height of the radar antenna determines the maximum range over which the radar can see. When a radar antenna is located 50 feet above the surface of the earth, the distance of the radar horizon is about 10 miles. If the radar antenna is 200 feet above the earth, the distance to the horizon is approximately 200 miles. When radar equipment is airborne, the range of detection is limited only by the power capabilities of the equipment because the distance to the horizon is practically infinite.

ANTENNAS

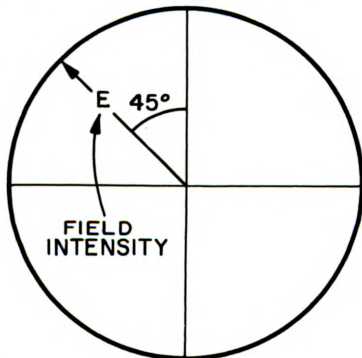
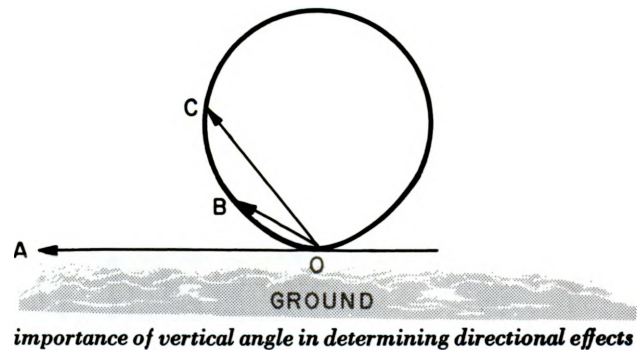
The antenna system of a radar set comprises the transmission lines or wave guides from the transmitter to the antenna and from the antenna to the receiver, any switching or protective devices in these circuits, the actual antenna or feed device terminating the transmission lines, and any reflector or lens that concentrates the electromagnetic energy into the desired

radiated beam pattern. In almost all radar systems the directional feature of the antenna system is most important; it makes possible the illumination of a specific target area and the reception of reflected signals from a specific direction. This directivity of the antenna is a measure of the overall efficiency of the system.

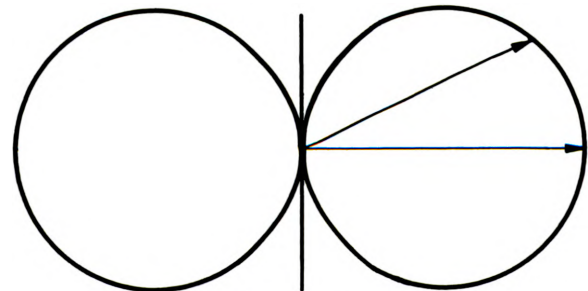
radiation patterns

The radiation pattern of an antenna is determined by making a polar plot of the electric field intensity measured at a given radius about the antenna while it is transmitting. The directivity function is the mathematical expression that relates the beam intensity to direction and is usually chosen to have a maximum value of one. Generally the radiation from an antenna can be fully defined by giving its radiation pattern in two planes, one containing the current-carrying elements and the other perpendicular to such elements. The radiation patterns of an elementary dipole are illustrated. Any linear antenna can be considered made up of a large number of such elementary current-carrying dipoles, connected in series. The radiation pattern of a receiving antenna is identical to

that of the transmitting antenna. This makes it possible for tests such as radiation pattern measurements to be made with the antenna receiving or transmitting.



the horizontal pattern of a vertical dipole antenna



the radiation pattern of a vertical half-wave antenna

methods of obtaining directivity

There are two basic methods of concentrating the radar energy to obtain directivity. The first method is to arrange two or more simple half-wave antennas (dipoles) so that their fields add in a specific direction and cancel or negate in other directions. Such a set of elements constitutes a linear antenna array, of which the common types are the yagi (or parasitic), broadside, end fire, and collinear. The second method is to use quasi-optical type antennas. Radar microwaves obey the laws of optics to a large degree and travel virtually in straight lines. Therefore in the quasi-optical types, the output of a simple feed horn or dipole

any may be concentrated by a parabolic reflector or by a lens system. Radar energy can also be beamed directly from horn type radiators, although in practice horns are used primarily to feed reflectors or lenses. Feeder horns are generally used with wave guide type feeds, while dipoles are used primarily with transmission line feeds.

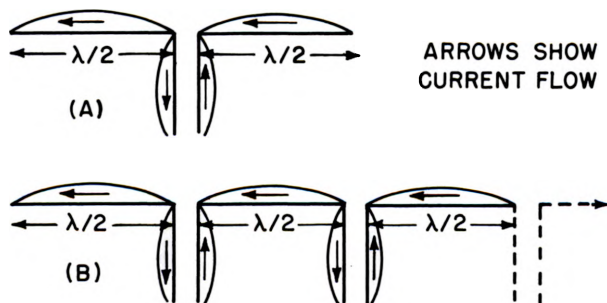
Most missile guidance antenna systems are of the quasi-optical type. In this type, the beam pattern from the feeder (horn or dipole plus small reflector) is called the primary pattern, and the beam pattern from the main reflector or lens is called the secondary pattern.

multielement linear arrays

The basic radiating element in a radar antenna is the center-fed, half-wave dipole, whose radiation pattern is similar to but narrower than that for the elemental dipole. Two or more dipoles can be arranged to produce an antenna array whose directivity function is the resultant of the fields of each element. The array pattern is a function not only of the element spacing but of the phase relationships of their feeding. The illustration shows how the array pattern for only two dipoles changes with spacing and phasing. Increasing the number of elements generally increases the directivity, as shown for a four-element linear array. The same directivity as with the parallel linear array can be obtained with collinear arrays.

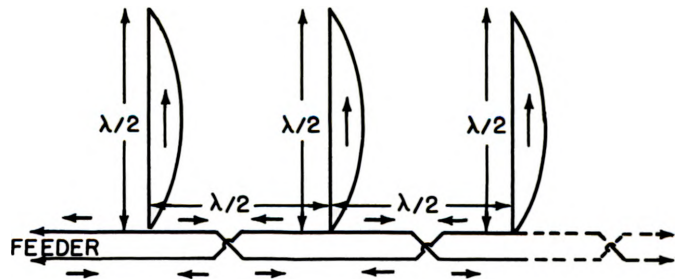
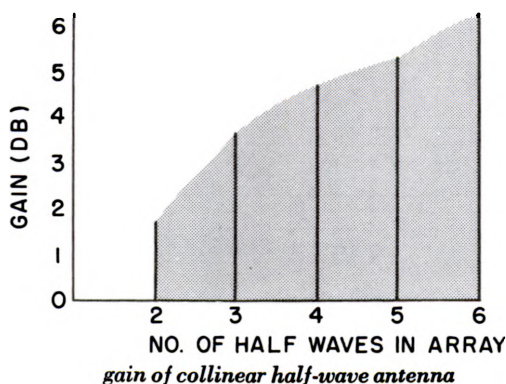
Very sharp antenna patterns in both horizontal and vertical planes with minimum of side lobes can be obtained by combining parallel and collinear array elements into a multielement broadside array. A screen or reflector placed one-quarter wavelength behind the array will give the broadside array an unidirectional radiation pattern.

These element arrays have fixed phase feeds. By arranging the phase and amplitude of the feed of the individual elements in a large multielement array, the shape and direction of the main beam can be varied. This principle is utilized in electronic scanning.

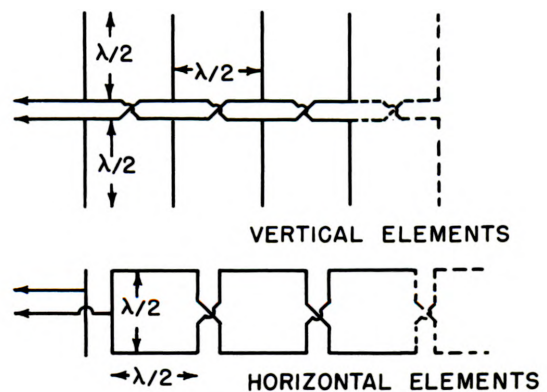
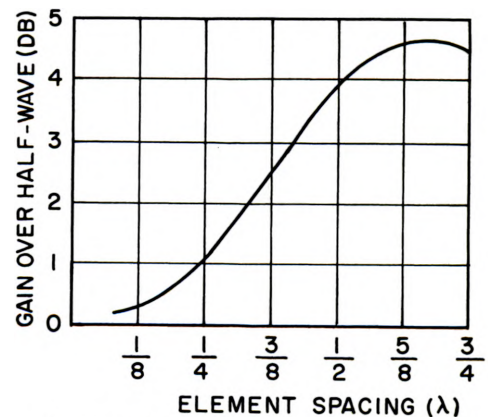
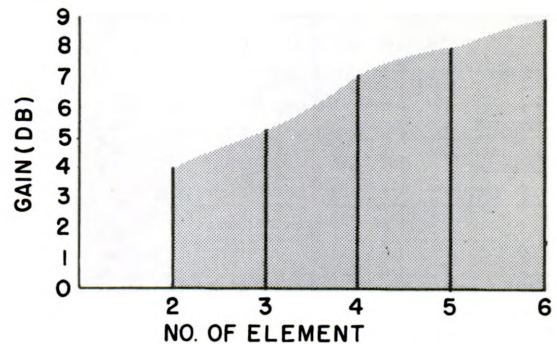


collinear half-wave antennas in phase

The system at A is generally known as "two half-waves in phase". B is an extension of the system; in theory the number of elements may be carried on indefinitely, but practical considerations usually limit the elements to four.



broadside array of parallel half-wave elements



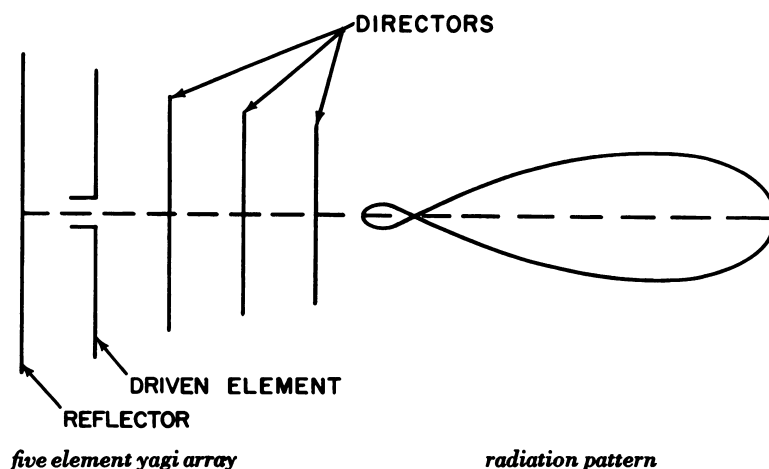
combination broadside and collinear arrays

Both arrays give low-angle radiation. Two or more sections may be used. The gain in db. will be equal, approximately, to the sum of the gain for one set of broadside elements plus the gain of one set of collinear elements.

PARASITIC OR "YAGI" ARRAYS

A driven antenna element is one whose energy input is furnished primarily by the output power of the transmitter. Conducting elements placed in the field of the radiating energy have voltages and currents induced in them. If the length of the nondriven conductor is predetermined at a half-wave length, these conductors can act as radiating and detecting elements on their own.

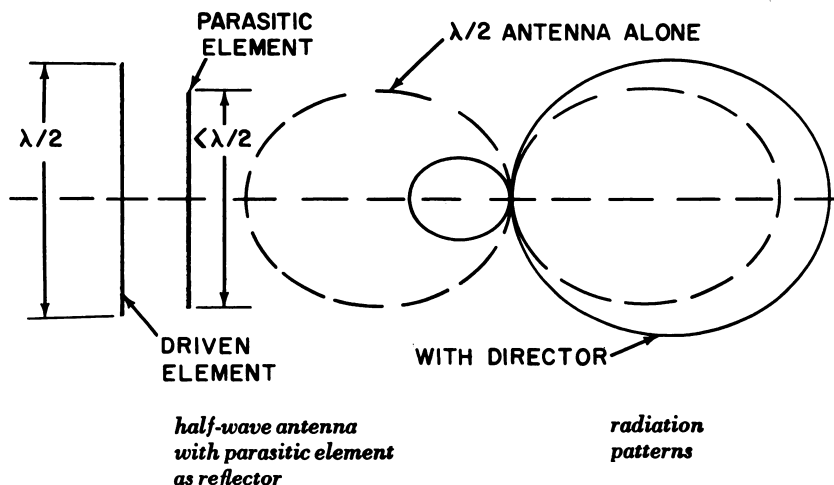
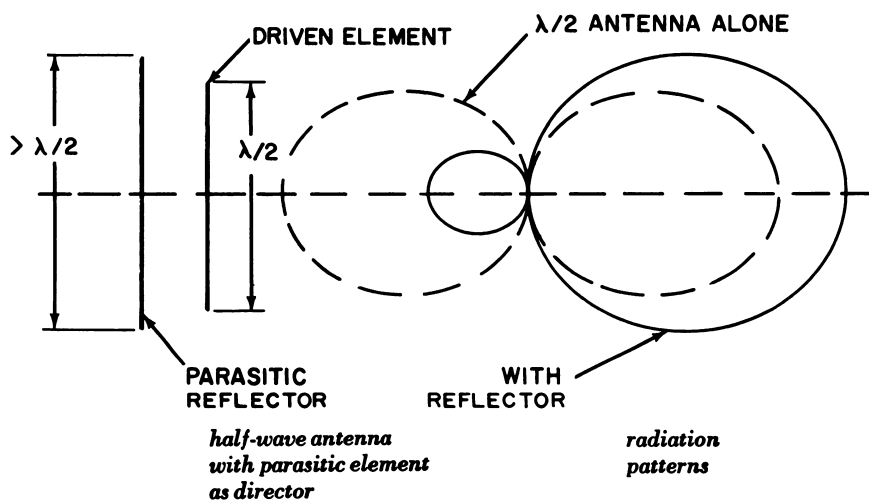
A one-half wavelength conductor in an antenna field will have voltages induced in it which will cause an antenna current flow in the conductor which in turn produces radiation. These nondriven elements are called parasitic elements. If such an element is placed near a driven element it will act as a reflector or director depending on length. A very highly directional antenna, known as a yagi antenna, is made up of one driven element and a number of parasitic elements.



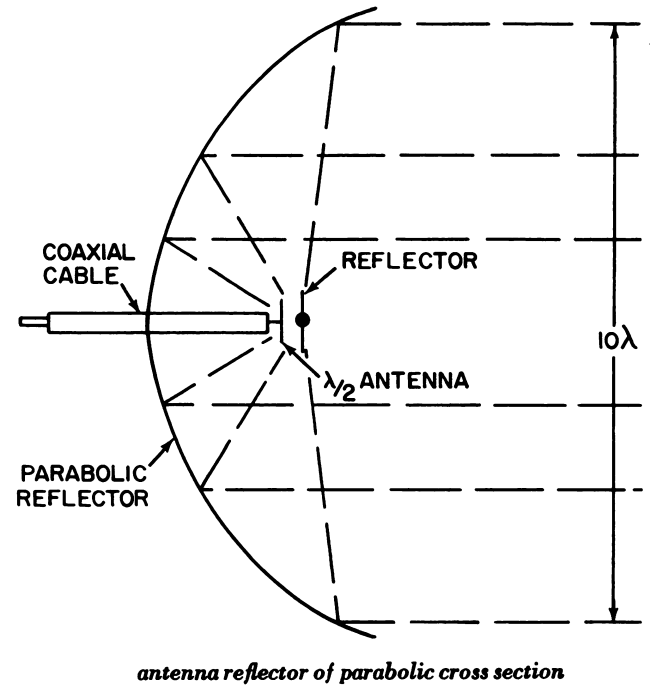
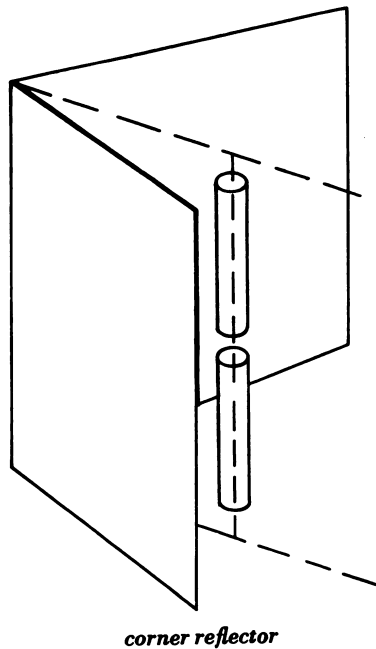
REFLECTORS

Various types and shapes of reflectors can be used with driven elements and feeders to obtain increased directivity and a unidirectional radiation pattern. In addition to reflecting elements and sheets, the other shapes commonly used are the corner reflector and parabolic reflector. At radar microwave frequencies, reflectors with parabolic cross-section are used primarily. These include the full paraboloid, the parabolic cylinder, the truncated paraboloid, and the orange peel paraboloid, in that order of usage.

The paraboloid or dish uses a point source, such as a horn or a dipole with a small auxiliary backup reflector as a feeder and produces a sharp parallel pencil beam.

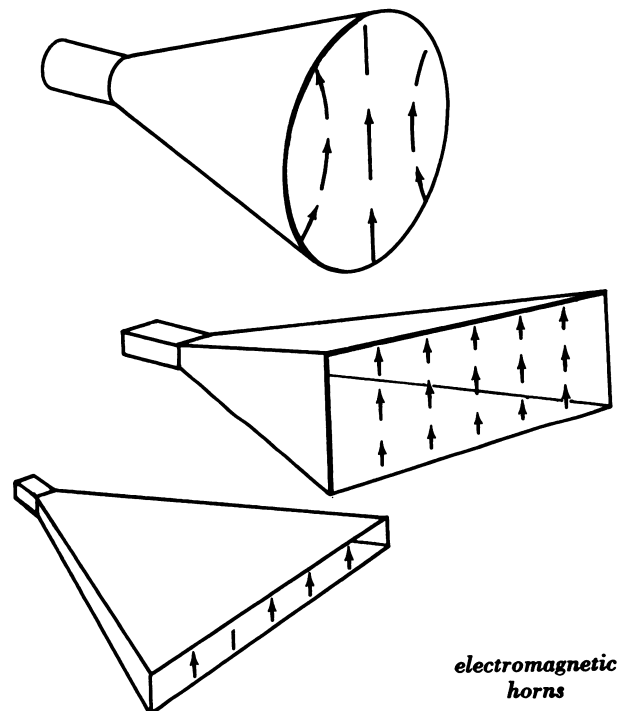


The parabolic cylinder uses a line source, such as collinear dipoles, and produces a flat fan-shaped beam to cover a greater volume of space for search purposes. The larger the reflector, the larger the effective antenna area and the sharper the beam. The reflectors are made of sheet metal or wire mesh.



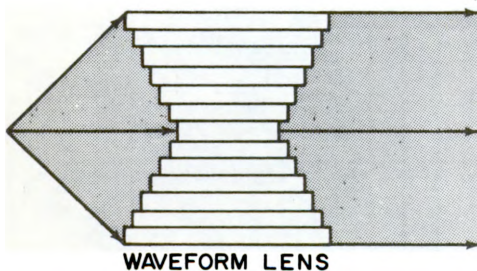
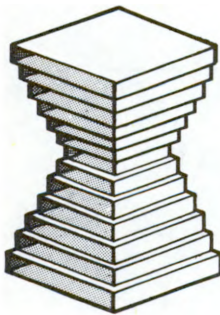
FEEDER HORNS

In high-frequency radar systems where wave guides are used as transmission lines, the wave guides can be flared into electromagnetic horns which can release and direct the waves into space. These horns are used generally as feeders to illuminate parabolic reflectors which produce the desired antenna pattern.

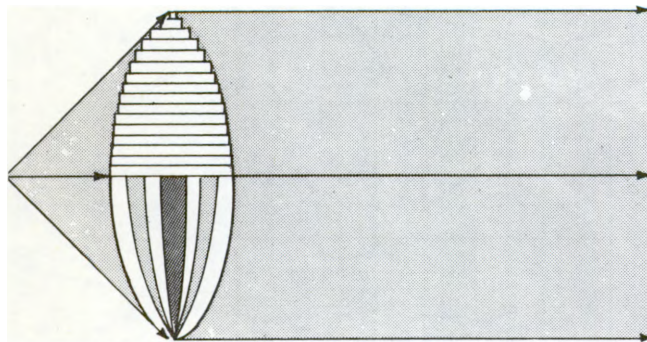


quasi-optical systems

In optics, lenses can be functionally interchanged with reflectors. Because microwaves behave similarly to light waves, lenses of various sorts can be used for micro-wave radar instead of reflectors or in conjunction with them. Some of the advantages of using lenses are: the feed horn or dipole can transmit directly through the lens towards the target instead of being in the path of a reflected beam, beam aberrations can be reduced, phase relationships can be better controlled, side lobes can be reduced, and electronic scanning is made easier because of the large area illuminated.



WAVEFORM LENS



DIELECTRIC LENS

radar lenses

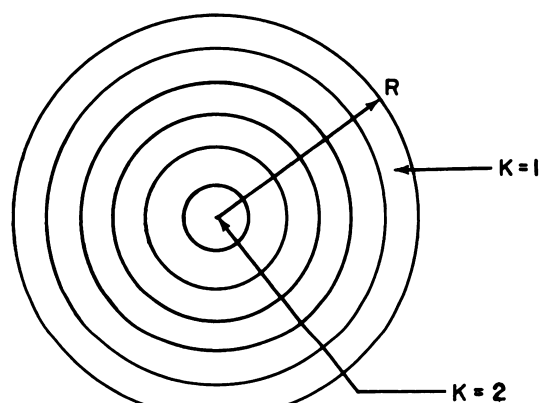
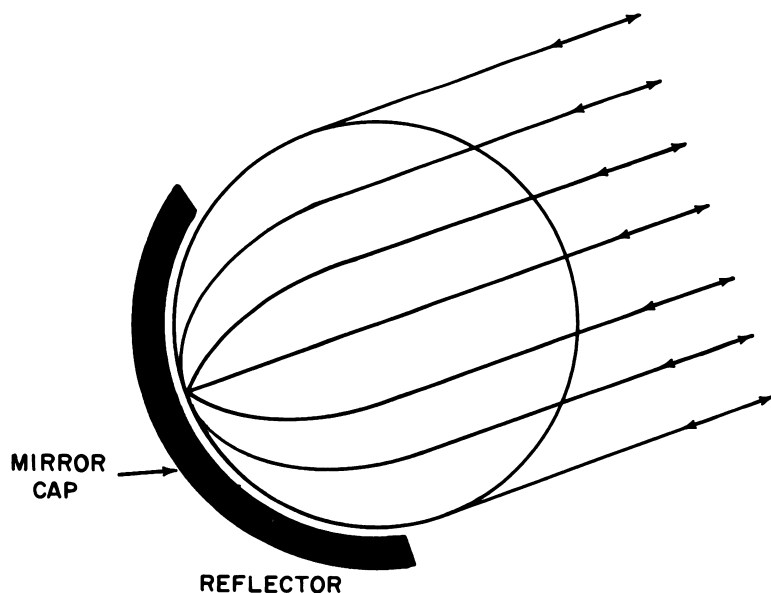
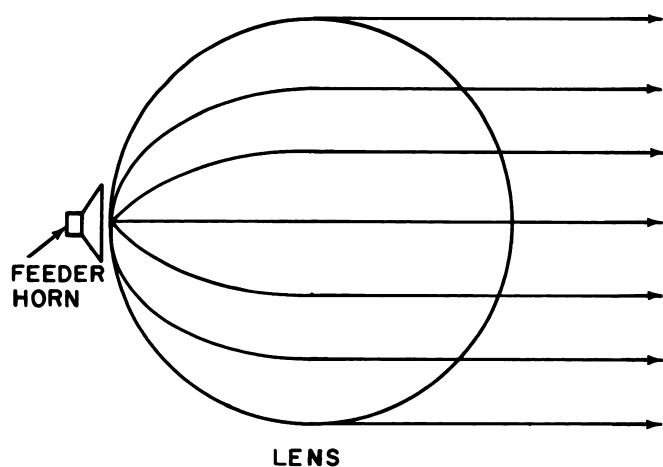
lenses

The purpose of any such lens is to refract the diverging beam from a feeder into a parallel beam, and vice versa. This can be accomplished theoretically by using as a lens material any medium in which the velocity of propagation varies from that in free space. For example, in a wave guide, the velocity is greater, while in a dielectric material, such as plastic with a dielectric constant greater than one, the velocity is less than that in air. Such a wave guide lens consists of a bundle of wave guides of varying lengths and looks like a concave egg crate. The outer rays of the beam are longer than the center ray. The outer wave guides are made longer to give the outer rays a longer high-velocity path so that the output of the lens is in phase in a plane perpendicular to the beam axis. In the dielectric lens, the short center rays are slowed down more in relation to the long outer rays to accomplish the same effect. Two forms of dielectric lens are illustrated. The upper half is a stack of rod-type lenses similar to the wave guide lenses. The lower half is fabricated from layers of material of different dielectric constants.

A special form of dielectric lens now under development is the spherical Luneberg lens. Ideally, a sphere in which the dielectric constant k varies with radius r as follows:

$$k = 2 - r^2,$$

has a focusing property such that all parallel incoming paths have the same electrical length from the incident tangent or phase front plane to the diametrically opposite part on the surface of the sphere. Likewise, the diverging beam from a feeder near the surface will be focused into a parallel beam out of the opposite side. This property is most useful for detection purposes. Since the sphere is symmetrical, moving a light feeder or sequentially energizing one of a series of feeders around the surface will move the beam in synchronism.



CONSTRUCTION METHOD

IDEAL RELATIONSHIP

$$K = 2 - r^2$$

Luneberg lens

reflectors

By backing up the sphere with a reflecting surface, a very efficient radar reflector is created with a larger radar cross section than the equivalent-size corner reflector. Any beam will be reflected back in the same direction from which it came over a wider angular range than is possible with the corner reflector. The Luneberg reflector will function up to a 180° conical angle. Since a material with continuously varying dielectric constant k has not yet been developed, the Luneberg lens is constructed in practice from a series of about 10 concentric plastic foam hemispherical half-shells whose dielectric constant varies in steps from 1 to 2. This lens is also known, therefore, as the stepped-index Luneberg lens. Another reflector lens combination used for radar is the cassagranian lens type. This was described in its application for IR systems and is essentially a folded optical system to conserve volume.

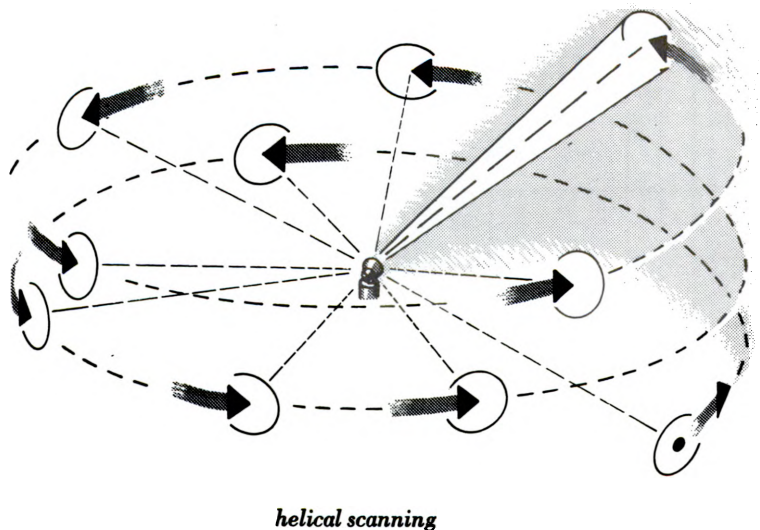
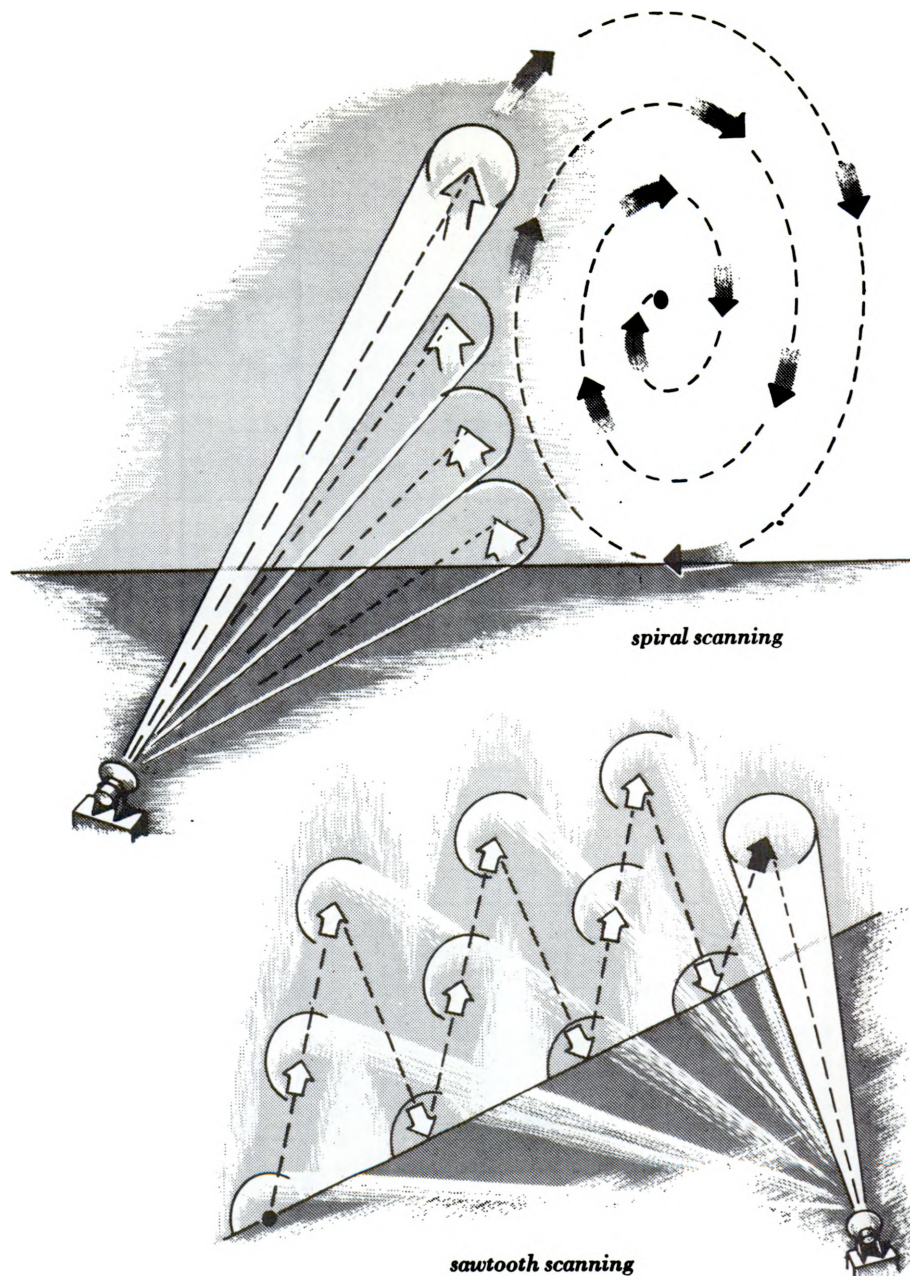
SCANNING

The systematic movement of a radar beam while searching for or tracking a target is referred to as scanning. The type and method of scanning used depends on the purpose and type of the radar and on the antenna size and design. The basic scanning procedures discussed here are also applicable to infra-red, light, sonar, or any other transmitting system that focuses a beam of energy.

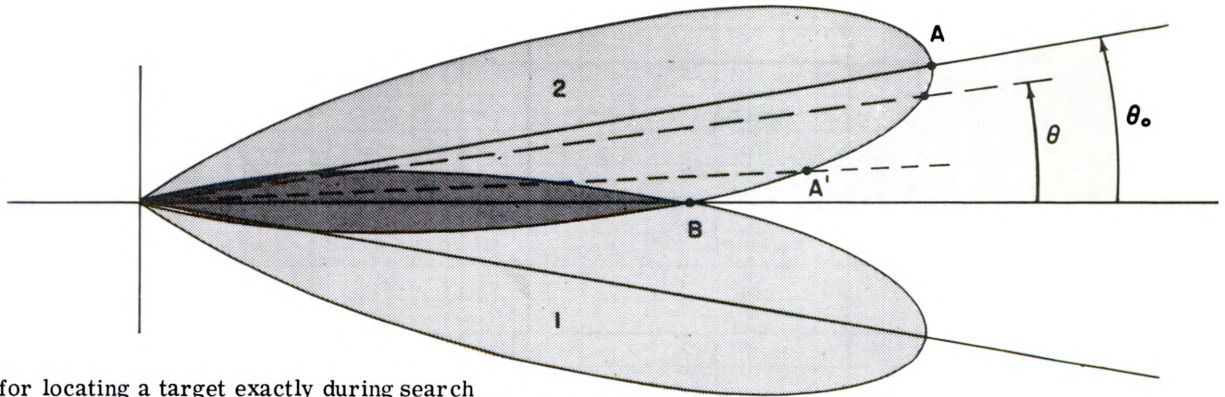
When searching for a target or for a missile that is to be guided, the whole area of interest must be scanned without gaps. This requires a rapid search carried out in a methodical manner. For a limited area search, such as when a target echo is lost because of fading or interference effects, a spiral or sawtooth scan can be used to cover the desired solid angle of space. For a complete 360-degree azimuth search, helical-type scanning can be used.

To insure solid coverage, the separation between paths of these scanning cycles must be approximately half the beam width so that there is some overlap. (As mentioned earlier, beam width is the angle between half-power points of the main lobe.) This means that there must be a fixed relationship between movements in different coordinates, i.e., in spiral scan between radial motion and circular rate, in sawtooth scan between azimuth and elevation rates, and in helical scan between azimuth and elevation rates. Broad searching patterns are achieved by simple azimuth rotation of an antenna with a fan-shaped beam where exact elevation information is not required.

The main considerations in such types of scanning are: the minimum scan period required for a given solid angle of search at the maximum range, the maximum angular velocity of the beam at maximum range to achieve a successful search, maximum target travel between successive scans in the radial direction at maximum range.



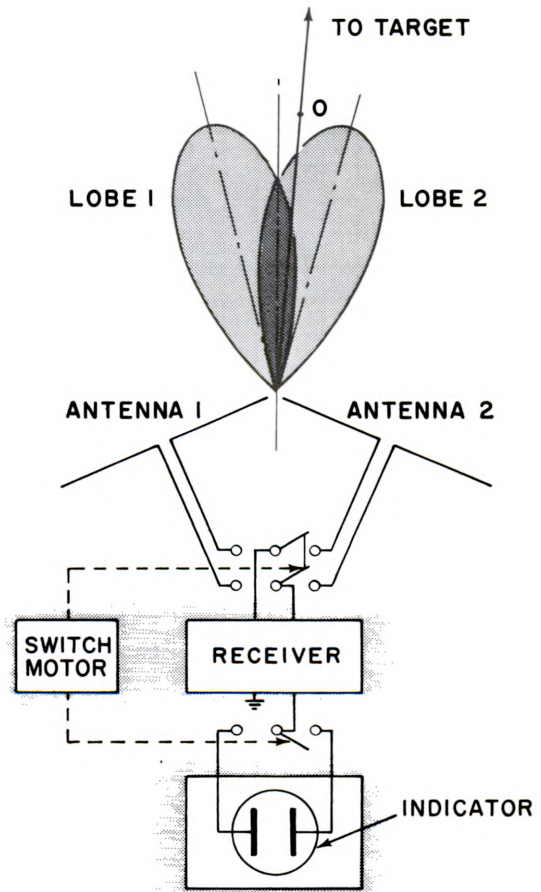
types of scanning



The procedure for locating a target exactly during search with a single lobe beam is to scan until an echo is received, then to point the antenna in the direction of the point from which a maximum signal is received. The accuracy with which this can be done is a function of beam sensitivity, which is the change in echo intensity for a given angular displacement in the plane of the beam. Referring to lobe 1 in the illustration, it can be seen that the sensitivity is a minimum at the beam axis and is generally a maximum between the 50% to 85% power points. It can be seen that the narrower the beam, the the greater the sensitivity.

However, the single lobe technique has certain drawbacks: there is the danger of losing a fast-moving target if the beam is too narrow; it is difficult to keep the beam axis exactly on target; there is no way of knowing which side of the beam an off-center target is on; and beam errors cannot be compensated for. These difficulties can be overcome by effectively using overlapping beams in one or more coordinates.

When overlapping beams are utilized to increase the accuracy of angular measurements, the effectiveness of the system depends on the sensitivity of the antenna to the received echo signals under optimum condition. For the greatest probability of successful target positioning, the axes of the two lobes must be displaced from each other sufficiently to cause the radiation patterns to cross over at less than the 85% maximum signal point. The advantage of the system is that the two signals are compared in magnitude and the operator is better able to rotate the antenna to an on-target position. An antenna system that employs the double lobe or overlapping lobe tracking procedure must include provisions for obtaining the two echo patterns and a method of comparing the intensities of the returned pulses. The simplest system would be to use two identical antenna systems, two receivers, a comparator, and one indicator. It is more economical however to use a lobe switching arrangement as illustrated. The antennas are positioned so that their patterns overlap, intersecting at approximately the half-power points.



mechanical lobe switch

methods of scanning

Two basic methods of accomplishing scanning are mechanical and electronic. In mechanical scanning, the beam can be moved in various ways. The entire antenna can be moved in the desired pattern; the feeder can be moved relative to a fixed reflector; or the reflector can be moved relative to a fixed source. In electronic scanning, the beam is effectively moved by such means as switching between a set of feeder sources, varying the phasing between elements in a multielement array, and comparing the amplitude and phase differences between signals received by a multielement array. A combination of mechanical and electronic scanning is also used in some antenna systems.

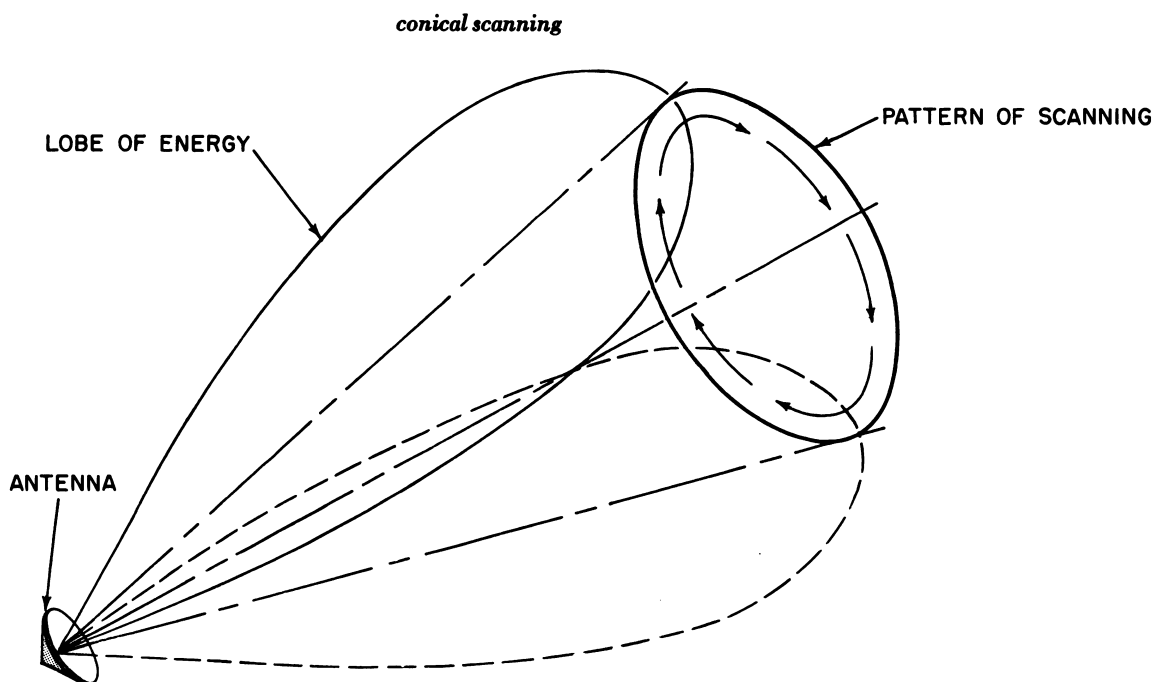
mechanical scanning

A common form of scanning for target tracking or missile beam rider systems is conical scanning. This is generally accomplished mechanically by nutating the feed point (moving the feed point) in a small circle around the focal point of a fixed parabolic reflector. If the feed is at the focus of the paraboloid, the antenna beam will lie along the axis of the reflector. If the feed point is moved transversely slightly away from the focus, the beam will be at an angle with the axis. If the feed point is oscillated back and forth, the beam will swing from side to side. If the feed point is moved (nutated) in a circle about the axis, a conical scan will result. A spiral scan can be created by slowly moving the feed point in and out along the longitudinal axis while rapidly rotating it, since the change in longitudinal position of an offset feed point will also change the beam angle.

With conical scanning, it can be seen that if the target is on the cone axis a constant amplitude signal is received, whereas if the target is off the axis the signal will vary sinusoidally with amplitude and phase depending on the angle and the direction away from the axis.

If the target is lost in a tracking radar with conical scan an axial motion can be added to the nutating feed, thus creating a spiral scan. When the target is again centered, the feed motion is changed back to simple conical scanning.

The disadvantages of mechanical or time-sequential scanning are: the amplitude-modulated target signals are accompanied and partially masked by external noise and the moving of an antenna involves mechanical problems. Among the mechanical problems developed are those created by the time lag in response to inertia effects of the moving parts and those developed by jitter effects on the beam resulting from vibration.



electronic scanning

Electronic scanning however, can accomplish lobe motion more rapidly and without the inherent mechanical disadvantages of conical systems. Because electronic scanning cannot generally cover as large a volume in space, however, it is sometimes combined with mechanical scanning in particular applications. There are two general methods of electronic scanning: lobe switching and monopulse or simultaneous lobing.

LOBE SWITCHING

In lobe switching, an electronic switch sequentially activates two or more feeder elements, each of which defines a different lobe.

The signals received from the different elements are then compared in amplitude and phase. This method generally requires long time constant storage circuits for integration of a large number of pulses to obtain sufficient data for comparison.

MONOPULSE OR SIMULTANEOUS LOBING

With monopulse or simultaneous lobing, all range and angle information is obtained from one pulse simultaneously received by a multiplicity of feeder elements which define adjacent and narrow fixed beams. If a target is on the center axis of the receiver antenna, the return echo to any one of the feeders arranged symmetrically about the center of the antenna will be in phase with the echo at any other feeder. If the target is off center, the echo will reach one feeder ahead of the other and the resulting signals will be out of phase. Proper phase comparison circuits can then establish direction and magnitude of target displacement.

Rate of scanning, is limited also by the allowable scanning losses for a given beam width. A scanning loss is seen as a reduction in maximum range from that obtainable with a fixed beam; the loss increases with scanning rate and decreases with beam width. The reduction in range results because during the time interval between the transmission and reception of an echo pulse the antenna has shifted to a point of lower gain on its pattern.

DISCRIMINATION AND RESOLUTION

When a target is to be tracked by radar some difficulty is experienced in separating the target echo from background noise. This difficulty is most severe when a fast target is to be acquired initially by a tracking beam, because the different action between target and echo is limited by the skill of the operator. During the subsequent tracking period, this difficulty remains acute in that the radar beam may not be locked on the desired target. The tracking of a single target is made difficult when there are other targets in the vicinity of the desired target. The ability of a radar or guided missile to separate a single target among a number of targets is called its resolving power. A radar set guided missile must be able to separate a target in an attacking formation if a missile is to be directed successfully to that particular target. A formation of attacking aircraft composed of one priority target and a number of decoys, for example, may be so constituted and flown that a tracking beam will be unable to fix on one aircraft; instead it will tend to drift from one target to another and will center eventually on a point midway between two targets. The effectiveness of a radar in defeating a countermeasure of this type will be dependent in general on the widths of its range and velocity gates relative to the spacing of

the aircraft in the formation. Although the complex and variable nature of external noise, which cannot be separated from the internal noise in the receiver, increases the problem of resolution of multiple targets; the utilization of a narrow-band velocity gate, a distinguishing feature of CW-Doppler radars, is effective in eliminating some of this external noise that is present with the target echo.

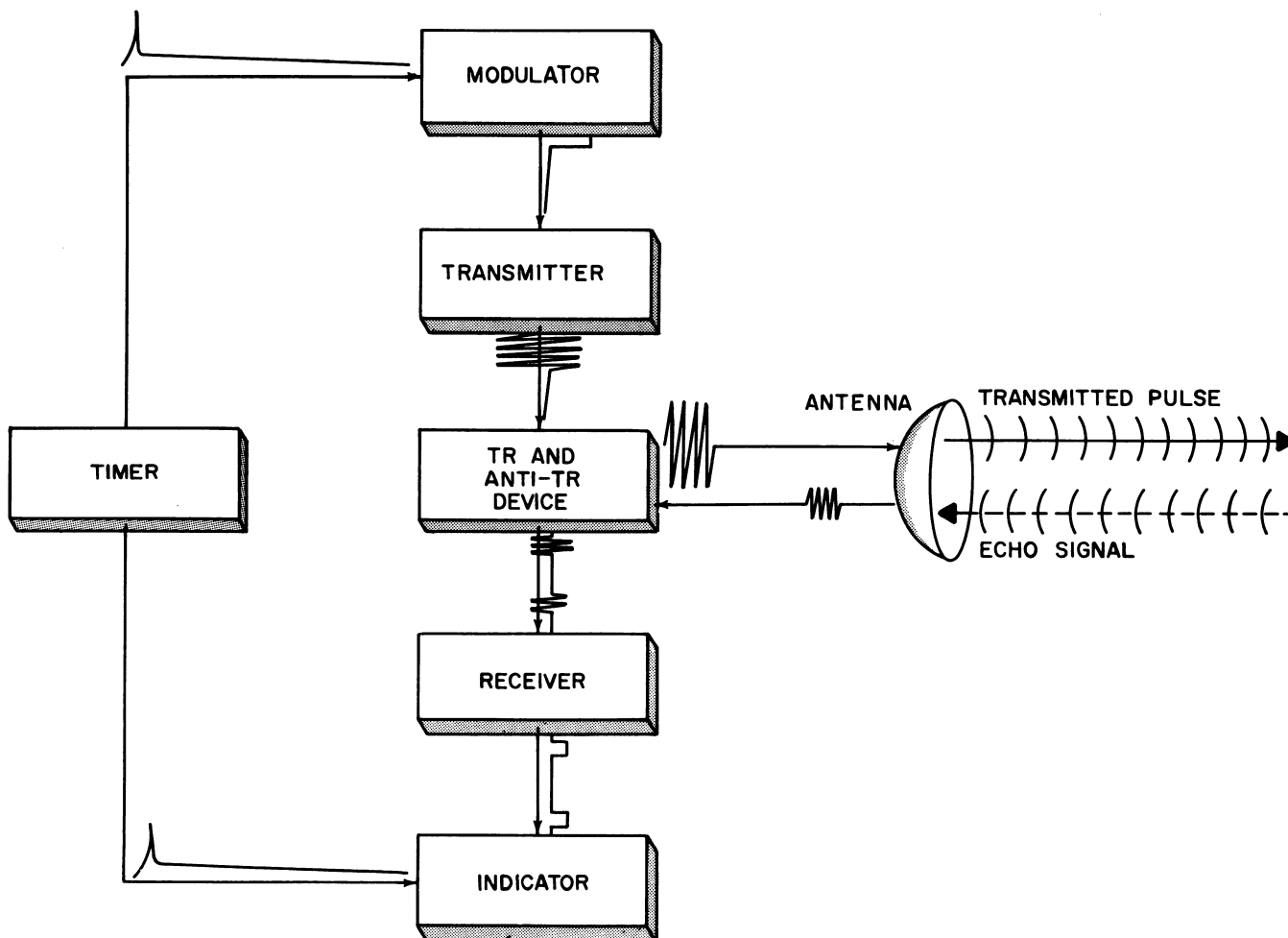
Range and velocity gating, used singly or together, permit the resolution in range and velocity of targets that are moving in radial paths toward the radar set. The ability of a radar to separate multiple targets which are at the same range but which are separated in angle depends on the optical resolving power of the beam. This factor varies directly with the wavelength and inversely with the effective diameter of the antenna.

A narrow beam is advantageous from the standpoint of noise discrimination, because it is less likely to pick up extraneous noise than a beam of wider angle. However, the improvement in angular resolution is usually at the expense of range capability; hence the advantage of high angular resolution must be compared with the possible disadvantages to the overall performance of the radar or guidance system.

FUNDAMENTAL ELEMENTS OF

Radar systems vary greatly in detail, but the basic features are essentially the same for all pulse radar sets. The functional block diagram shown includes the basic components that comprise a pulse radar system. The functional breakdown resolves itself into seven essential components:

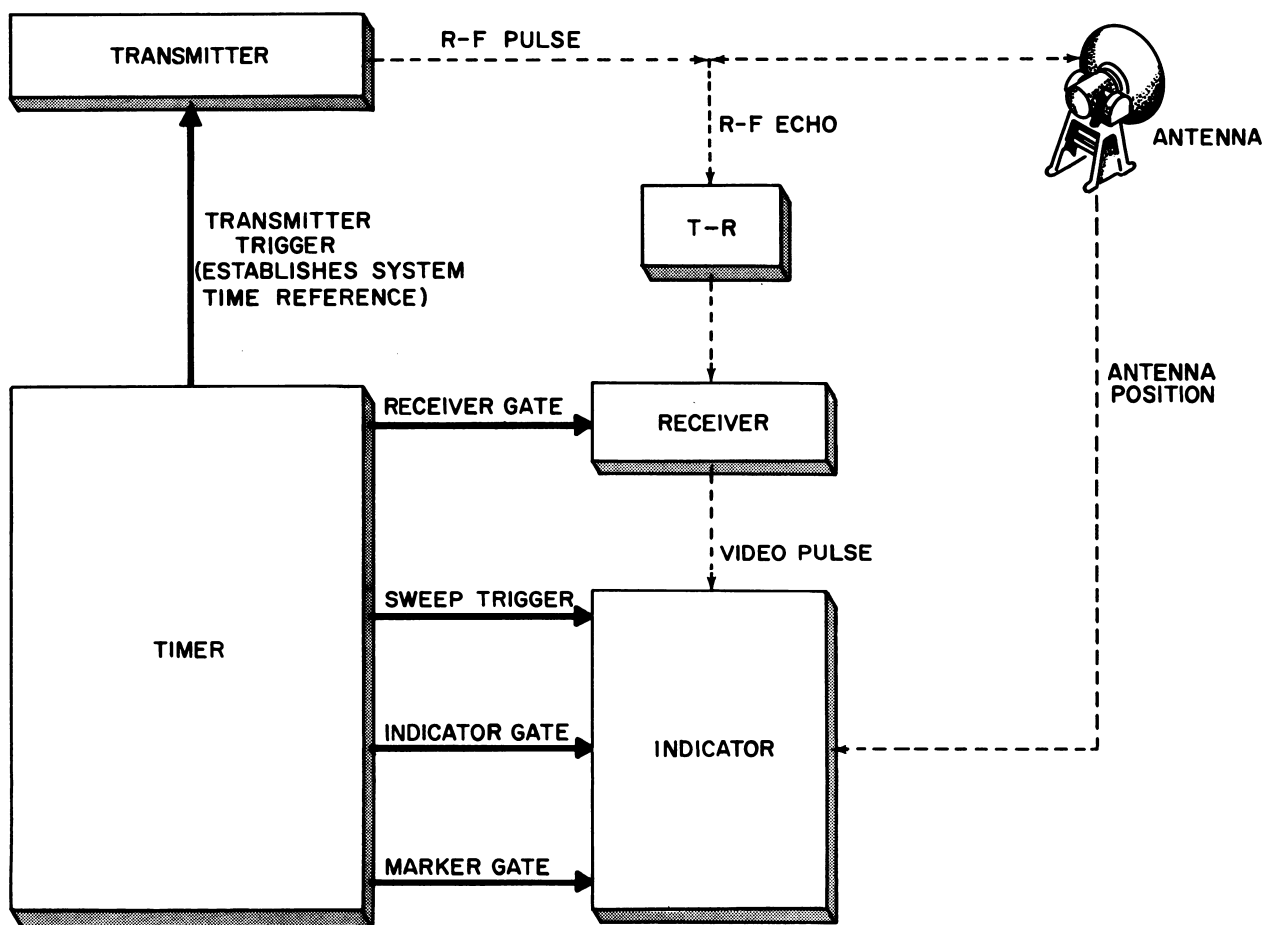
- | | | | |
|------------------------|---|-------------------------------|---|
| (1) The timer | — synchronizes the transmitter pulse with the initiation of the time base in the indicator unit | (5) The TR and anti-TR switch | — Most radar sets employ a single antenna for both transmission and reception. This device assures that the receiver presents a high impedance path to the transmitted energy during periods in which the radar set is transmitting and that this impedance is eliminated during periods in which the set is receiving. It is commonly known as the TR (transmit-receive) switch or as a duplexer, and in certain instances as a polyplexer |
| (2) The transmitter | — generates RF energy in the form of high-power pulses | (6) The receiver | — amplifies and detects echo pulses, and produces amplified output video pulses sent to the indicator |
| (3) The modulator | — provides a high-voltage pulse which modulates the transmitter and forms the transmitted pulse. | (7) The indicator | — provides the visual display of radar information. |
| (4) The antenna system | — the antenna has two prime purposes:
1) radiates the RF energy output of the transmitter in a highly directional beam
2) Detects or receives returning echo energy and forwards it to the receiver | | |



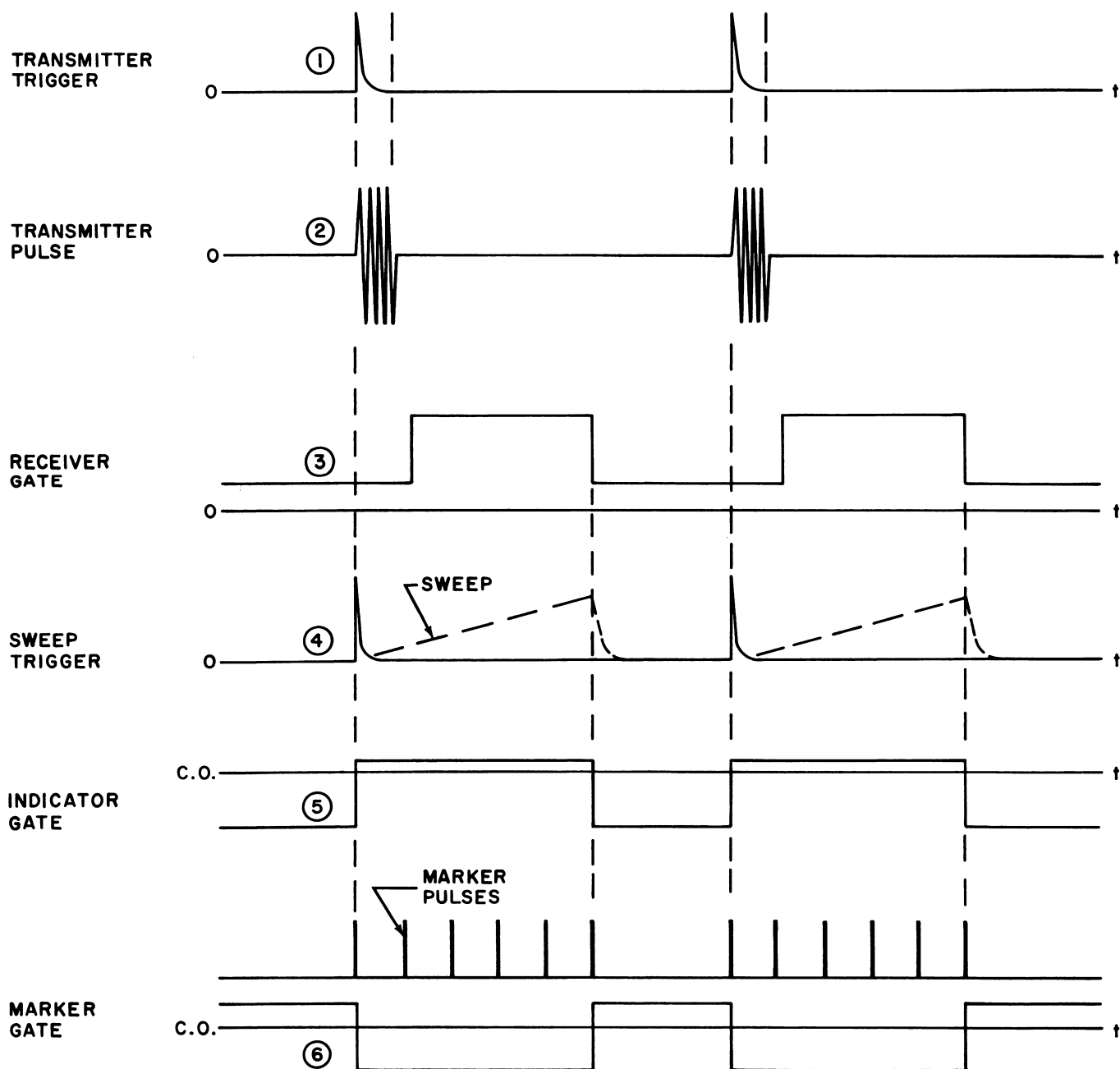
PULSE RADAR SYSTEMS

the timer

The timer performs the functions of establishing the pulse reflection rate of the radar system and of insuring that the modulator and the indicator operate in a definite time relationship with each other and initiate their functions at the instant that the transmitter produces a pulse of RF energy. Most radars are timed or synchronized by a self-blocking oscillator, a sinusoidal oscillator, or a multivibrator, or by the rotary mechanical motion of a spark gap modulator. When oscillators are used to control the repetition frequency, the oscillator output is shaped or sharpened by peaking circuits before being applied as a trigger pulse to the transmitter to synchronize the start of the indicator sweep. The master timing signals may be taken from a ship master oscillator, in which case the radar is said to be externally synchronized, and the master timer in the radar is turned off or bypassed.



As illustrated, timing triggers are used for synchronizing and gating, that is for turning circuits on and off at the proper time. The timing triggers, which may be of positive or negative potential, are the output pulses developed in the timer network. The timing triggers that gate the receiver are usually sent through a delay device so that the receiver does not become operative until a short time after the pulse is transmitted, preventing receiver loading during pulse transmission. Timing triggers are used to gate both the indicator and the marker circuits that may be located in the indicator unit. A trigger is used to start the sweep function of the indicator at the proper time.

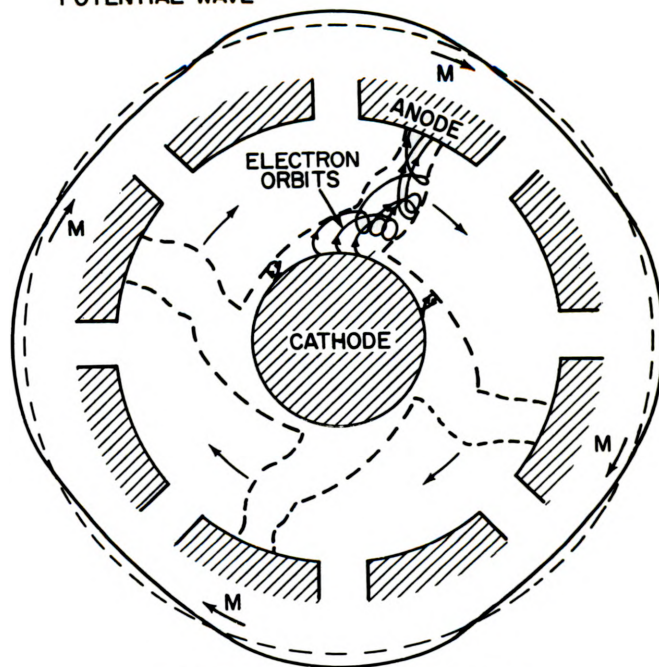


the transmitter

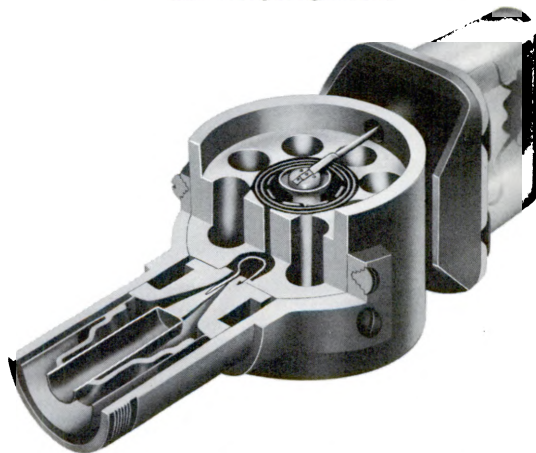
The transmitter functions to generate high-power pulses under the control of the timer. The high-frequency oscillator used in most radar transmitters is a resonant cavity magnetron type. The frequency generated by the magnetron determines the wavelength of the radio waves that are radiated into space. The higher the frequency, the shorter the wavelength. Since short wavelengths can be concentrated into a beam by a relatively small antenna, the wavelengths utilized in military applications are only a few centimeters long, and often of an order of a centimeter or less. In order to produce such short wavelengths, the high-frequency oscillator must generate voltages that alternate at a frequency of several billion cycles per second. The magnetron is well adapted to this high-speed task.

A magnetron consists primarily of a block of copper with resonant cavities inside. The magnetron is usually a diode with a cylindrical anode and an axial, filament-type cathode. The strong magnetic field from an external magnet is directed parallel to the filament, and when a pulse voltage from the modulator excites the cathode, electrons are caused to flow from the anode cavities by the action of the magnetic field. The cavities are capable of sustaining a definite frequency, called the resonant frequency. The size of the cavity determines the frequency that will be sustained. The smaller the cavity, the higher the frequency. A typical microwave cavity magnetron is shown. The power output is obtained either by a coupling loop or by a transformer section of a wave guide inserted into a cavity. The loop conveys the power generated inside the magnetron to the wave guide and from there to the antenna to be sent out into space.

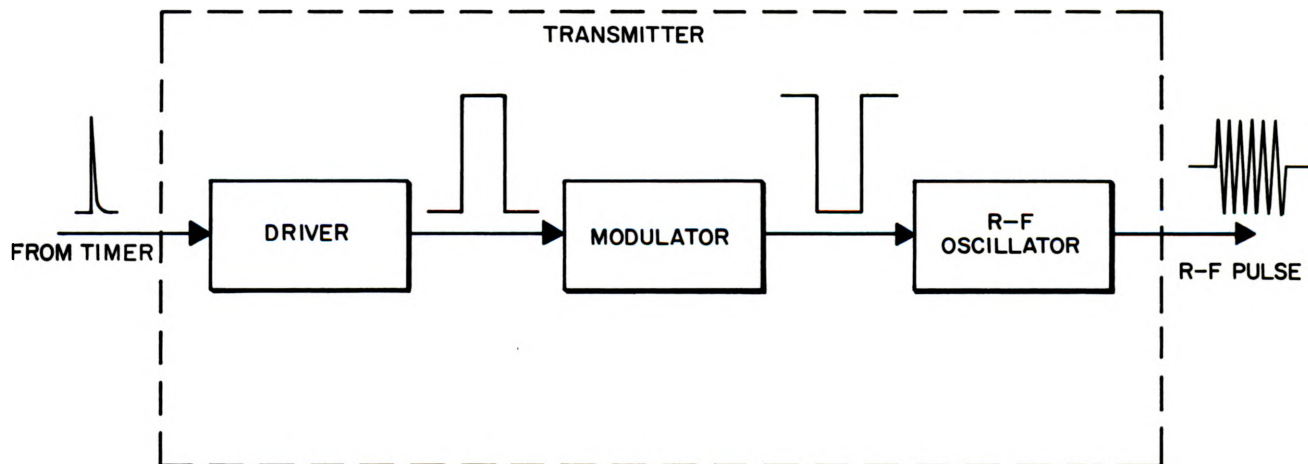
ROTATING ANODE
POTENTIAL WAVE



orbits of electrons of different phases
in a cavity magnetron



cut-away view of a cavity magnetron



externally pulsed transmitter

the modulator

Since the magnetron is a self-excited oscillator, it is necessary for the modulator or pulse generator to modulate the full output of the tube. The term modulator is used to refer to the circuits which control the application of high voltage to the magnetron. Since the expression modulate means to shape as well as to control, the modulator refers to the pulse-forming network as well as to the driver circuitry which produces the pulse that is delivered to the control grid of the modulator stage.

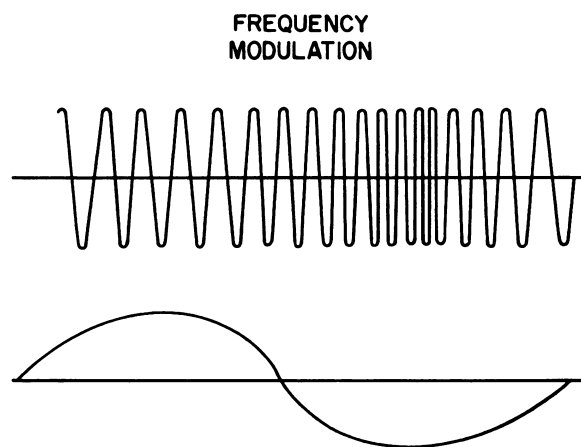
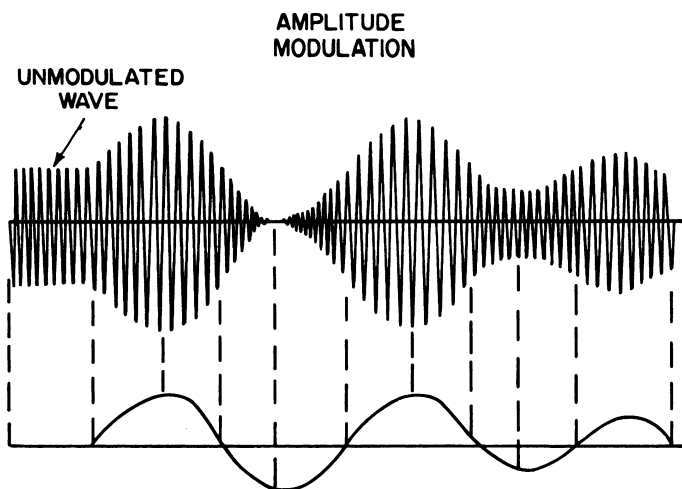
Modulators are also called pulsers or keyers to denote their circuit function. There are two basic types of pulsers, hard-tube and line-type pulsers. The term hard-tube refers to a vacuum tube which controls a capacitor discharge. The pulse is formed in a driver unit and applied to the control grid of the vacuum tube which acts as a switch to control the pulse duration.

In line-type pulsers, a lumped-constant transmission line serves as both the energy storage driver and the pulse-shaping element and is called a pulse-forming network (PFN). Whichever type is employed, the modulator furnishes a high-value voltage to the RF oscillator for the predetermined pulsing time, and is responsible for the pulse duration time of the RF oscillator. The outstanding characteristic of micro-wave magnetrons is the high power output that can be obtained when they are pulsed. The pulse power output often is more than one thousand times higher than a CW output at the same frequency.

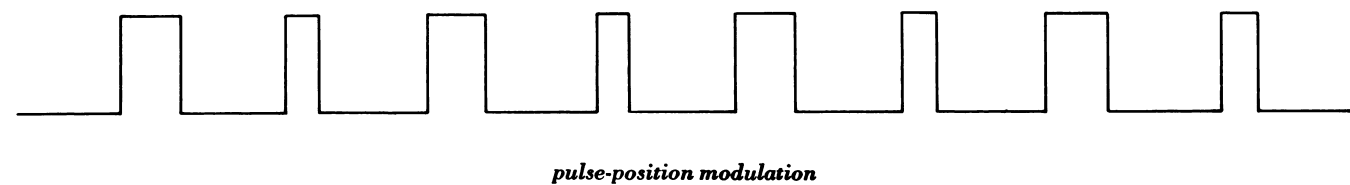
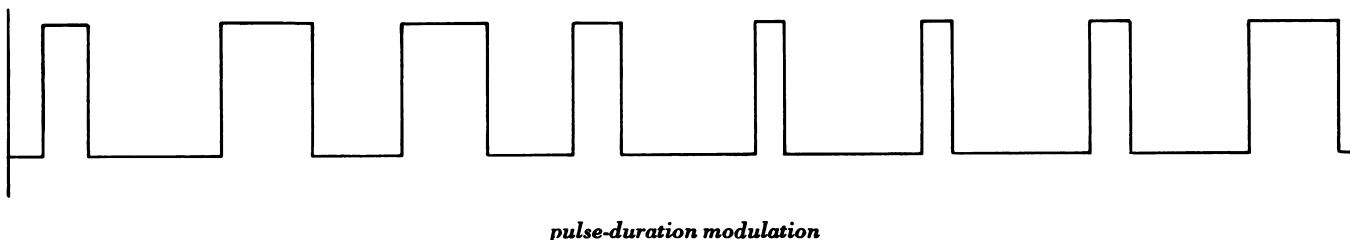
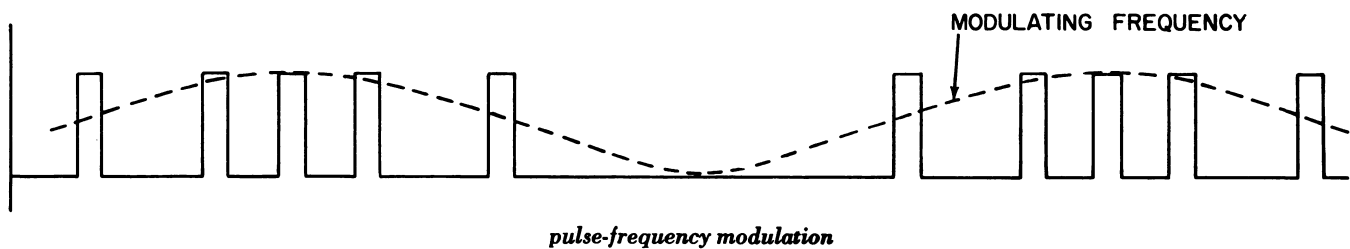
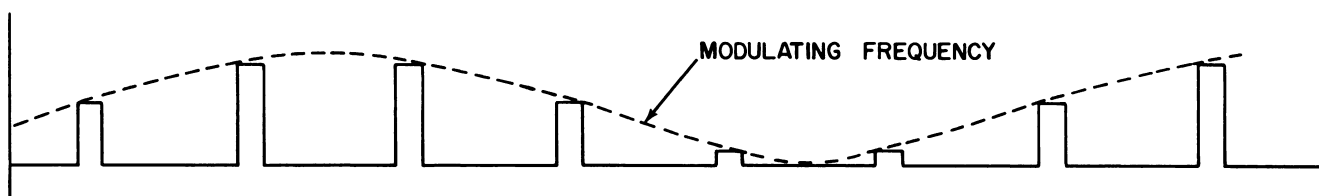
types of modulation

Modulation is the process whereby the carrier may be altered in any of three fundamental ways: by varying the amplitude of the carrier (amplitude modulation), by varying the frequency of the carrier (frequency modulation), or by varying the phase of the carrier (phase modulation). Variation of each of the modulating procedures may be superimposed on the pulsed output of a radar wave. When the amplitude of the pulsed radar signal is modulated (PAM) a direct means of controlling the range of a radar system is provided because the output power is directly proportional to the square of the amplitude. Also changes in range can be measured directly, as the amplitude form of modulation does not create a phase shift of the signal.

Frequency modulation involves a shift or change in the frequency of the transmitted signal. If the frequency of the transmitted energy is varied continually over a specified band (PFM), then the frequency being radiated differs from the received frequency. This difference results because the original frequency travels to the target and returns. Because the frequency difference depends on the distance traveled, it can be used as a measure of range. In pulse-pulse modulation, the duration of the pulse is varied (PDM) by changing the time constant of the modulating voltage pulse. Any modification of a pulse parameter with time is called pulse-time modulation (PTM). Circuitry for this type of modulation, although difficult to design, is an effective deterrent to enemy countermeasures. Another effective countermeasure tactic utilizes a coding arrangement of the pulsed output to identify a reflected signal as actually emanating from a target, and not an enemy countermeasure device. The simplest codes and the most effective are the binary and ternary types. Computer inputs are designed to accept binary information at almost limitless speed. Also more elaborate codes are possible and may involve a continuously varying function. Multiplexing (the transmission of multiple messages on a radar pulse) can be accomplished by a time or frequency division of the pulse signal. The instantaneous amplitude of the signal is sampled, one channel at a time, and transmitted in regular time sequence until all channels (pulses) have been sampled. The nature of the modulation determines the pulse width (duration). Since only one instantaneous value is transmitted at one time, there is no inter-channel modulation. When pulse modulation is combined with time division the system is known as a pulse-time-multiplex system.



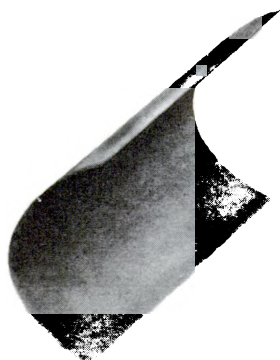
modulated waves



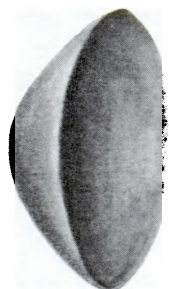
the antenna system

The complete antenna system includes the antenna assemblies (energized elements and reflectors), the antenna positioning system, and the lobe-switching or conical scanning system. The energized elements illuminate the reflectors, which are designed to form the desired beam pattern. The most common reflectors are parabolic dishes (paraboloids), truncated para-

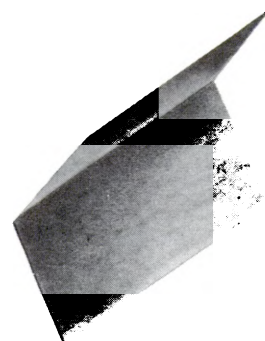
boloids, and corner reflectors. The shape of the antenna pattern is dictated by tactical considerations. Surface-search radars can employ a narrow beam width in elevation and a wider beam in azimuth. This gives better coverage in azimuth but less accuracy. Air search radar must have a beam form that extends from the zenith to the surface.



PARABOLIC CYLINDER



PARABOLOID

TRUNCATED
PARABOLOIDORANGE-PEEL
PARABOLOID

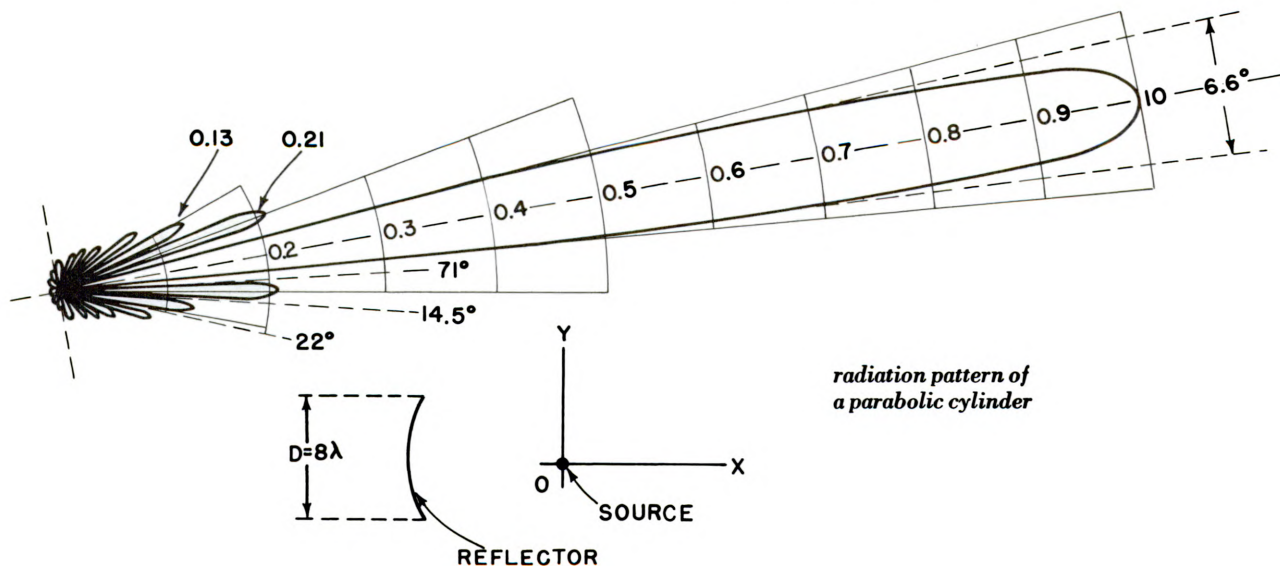
CORNER REFLECTOR

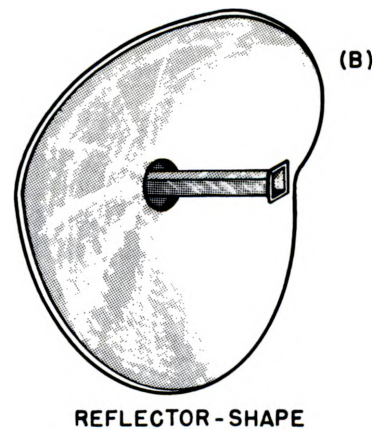
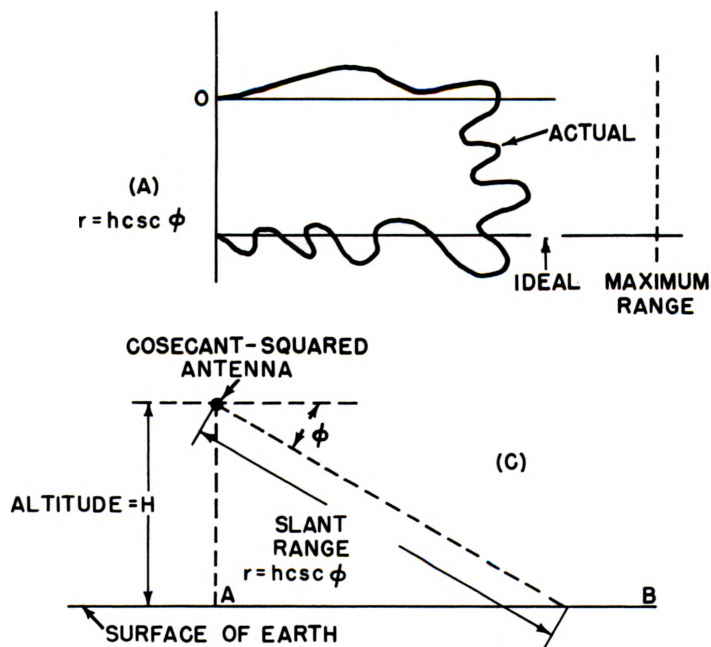
Fire control radars require a narrow pattern in both azimuth and elevation in order to obtain high resolution and high accuracy. The radiation pattern from a parabolic cylinder is shown.

The pattern is the pattern in the xy (vertical) plane of the cylinder, which has a width (D) of 8 wavelengths and an infinite length in the t direction, and which is illuminated uniformly by a line element. The mathematical analysis of the field strength at any point shows that the field is a function of $\sin x$ where x is a function of D ,

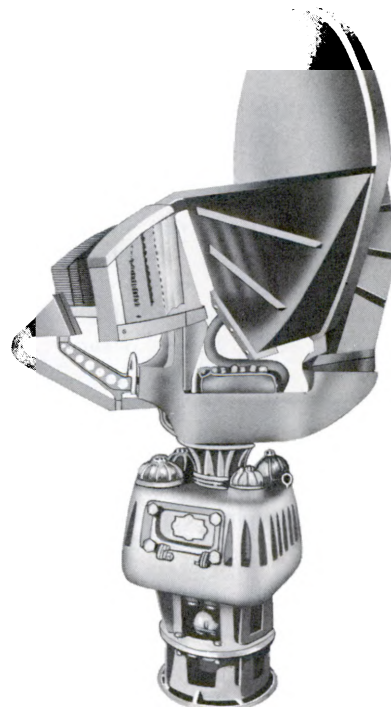
λ , and θ . The angle θ is the angle off center of the beam. The beamwidth, therefore, is found by the use of the function $\sin x/x$ and the pattern is referred to as a $\sin x/x$ pattern. The field strength from a paraboloid is a function of $\sin x/x$ in both azimuth and elevation. The beamwidth is narrowed by increasing the dimension D of the parabolic cylinder, or by increasing the diameter of the paraboloid dish.

A basic problem in antenna design is that all aircraft radar at equal altitudes must give equal target indications to assure exact identification of surface targets.

radiation pattern of
a parabolic cylinder



An antenna capable of solving this problem is the cosecant squared antenna. The cosecant squared antenna develops a beam pattern such that the power density varies as the square of the cosecant of the angle θ as shown. The slant range from an airborne antenna to a point on the surface of the earth is $h (\csc \theta)$, where h = height of antenna and θ = the depression angle. The antenna is designed so that the field intensity pattern also varies as the cosecant of θ . The airborne radar thus lays down a uniform electric field along a line of the earth's surface, and all targets on this line give equal indications. Since power density varies as the square of field intensity, the power density pattern varies as the square of the cosecant of θ , and hence the name cosecant squared antenna.



cosecant-squared antenna

r-f transmission line and wave guide

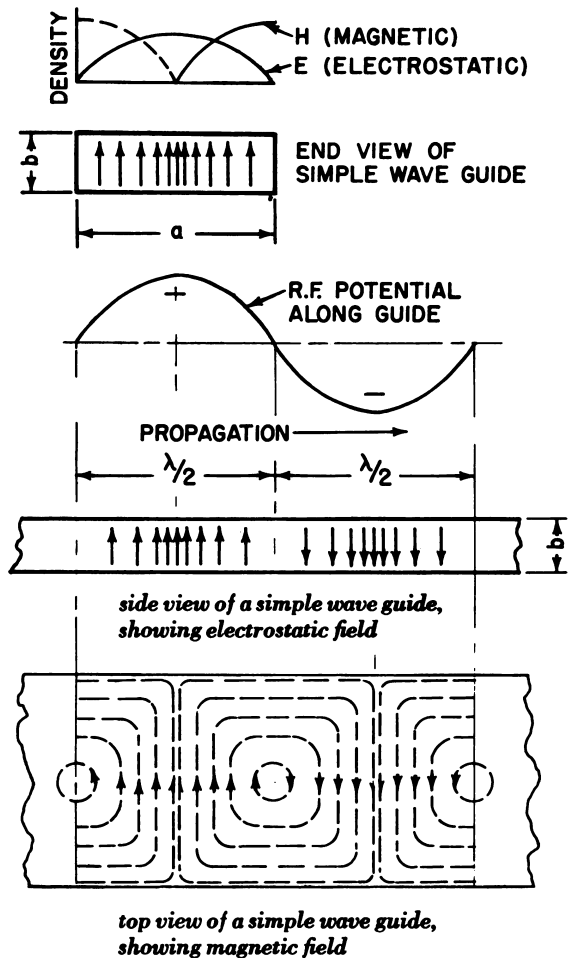
The RF transmission lines in naval radars are either coaxial lines or rectangular wave guides. In an optimum system, all of the output power will be transferred from the transmitter to the antenna. This efficiency can be approached only by matching the antenna impedance to the characteristic impedance of the wave guide. A wave guide is a simple hollow metal tube and may be rectangular, round, or oval in shape. The transmission of RF energy in wave guides is a function of the frequency and the distribution constants of the transmitted fields. Wave guides are essentially ultra-high-frequency transmission devices, since the frequency must be of a high order before a field can be successfully transmitted through a wave guide. A wave guide has a definite cut off frequency determined from the cross section of the tube, and will not operate at a frequency lower than its characteristic cut off frequency.

In dealing with transmission inside a hollow conductor we have to abandon the usual concepts of current and voltage, for in the absence of a second conductor there is no ordinary complete electrical path. It is logical to use the more fundamental conception of a field of force characterized by electric and magnetic vectors. For a simple rectangular wave guide, the electrostatic lines of force and density are shown in the illustration. The electromagnetic lines of force are perpendicular to the electrostatic lines of force traveling down the tube in the direction of propagation. A rectangular wave guide will transmit satisfactorily if the component of the electric field tangent to the side surface is zero at every point on the surface.

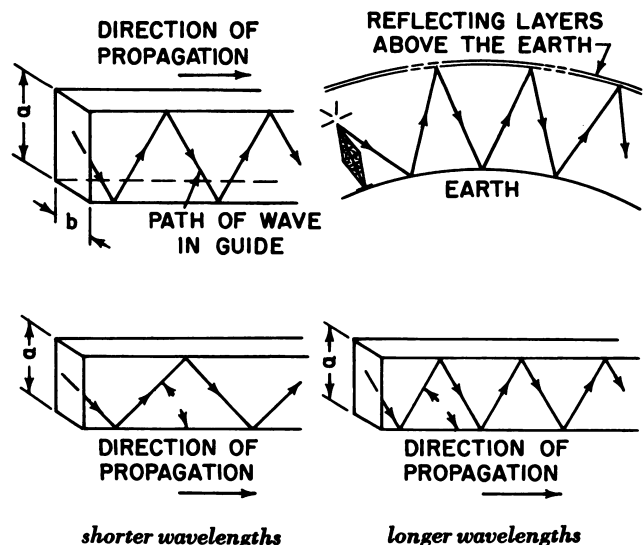
An analogous device is a speaking tube used with sound waves. Ordinarily, sound waves spread out in all directions, but are reflected from solid objects. Even giving sound directivity by use of megaphones does not make for efficient power transmission. However, a speaking tube made of sound reflecting walls will confine the sound to a path within the tube by having it bounce back and forth from wall to wall while traveling down the tube. In a similar manner, wave guides may be used to confine electromagnetic waves within the guide.

types of propagation

There are two basic types of wave propagation through a wave guide: TM and TE type propagation. TM propagation (transverse magnetic) has no magnet field (H) in the direction of propagation down the tube. The electrical field (E) has a longitudinal component in the direction of propagation. There are numerous modes of operation, depending upon the cross-sectional area of the tube and the method of RF excitation. Different modes are identified by subnumerals after the letters TM. In TE propagation (transverse-electrostatic), the magnetic field (H) is in the direction of propagation and the electric field (E) is transverse. There are various modes of operation and they are identified as mentioned in the paragraph above.



The velocity of propagation of RF energy in a wave guide is less than that in air because the wave does not travel in straight lines through the guide but is reflected from the inner wall. Velocity of the wave train depends, therefore, on the frequency and the tube dimensions. Wave guides are used for other purposes besides the transmission of power: they may be used as circuit elements and matching devices. They may also be used as measuring instruments and as rotating elements in antenna systems.

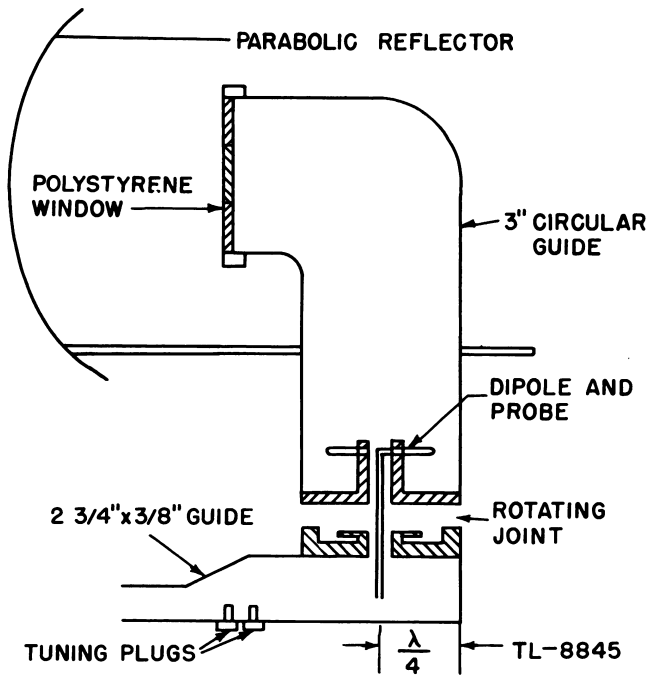


rotating choke joint

The RF lines in naval radars are either coaxial or rectangular wave guides. To deliver RF power from a transmitter power to a rotating antenna, a rotating transfer joint is necessary. The joints commonly used are of the noncontact choke type.

In a coaxial type joint, the gap between the stationary and rotating parts must be located at a zero impedance point (end of quarter-wave section of line) to allow power to flow across the line without sparking or loss. To prevent radiation from the open end of the choke joint, a short-circuited quarter-wave section is added in series. These sections of line are called chokes and hence the joints are called choke joints.

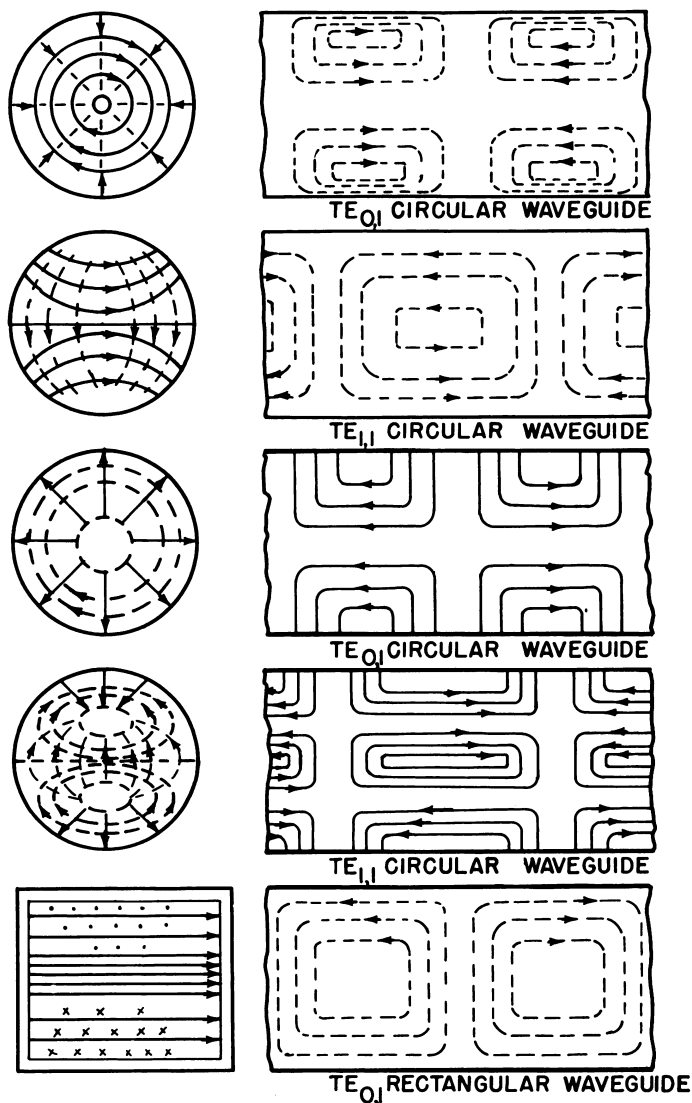
Rotary joints between wave guide sections employ the same basic principles but use a circular wave guide rotating joint. The same folded choke arrangement is employed to permit currents to flow but not to escape from the junction. Since the modes of propagation in circular wave guides are different from those in rectangular wave guides, the joints must be carefully designed to minimize coupling of undesired modes from one rectangular guide to the other, and various absorbers for the undesired modes are used.



coaxial rotating joint

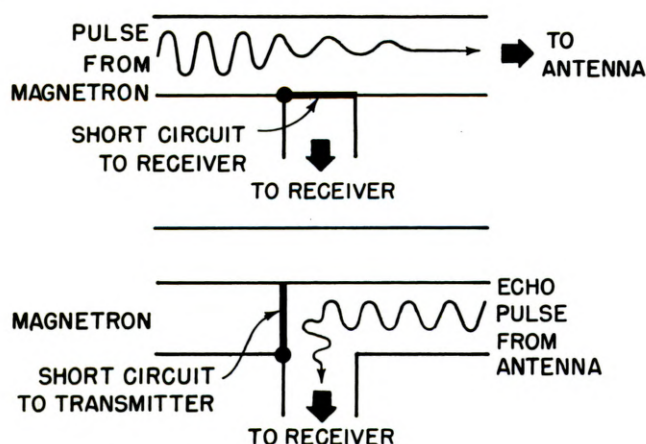
*common modes of
waveguide excitation*

The mode of excitation has considerable effect on the losses which occur and as a result there are three modes which are commonly used. Of the two circular modes the $TE_{0,1}$ has the greater losses, but it will maintain the polarization of the wave in passing through a rotating joint and therefore is used quite often.



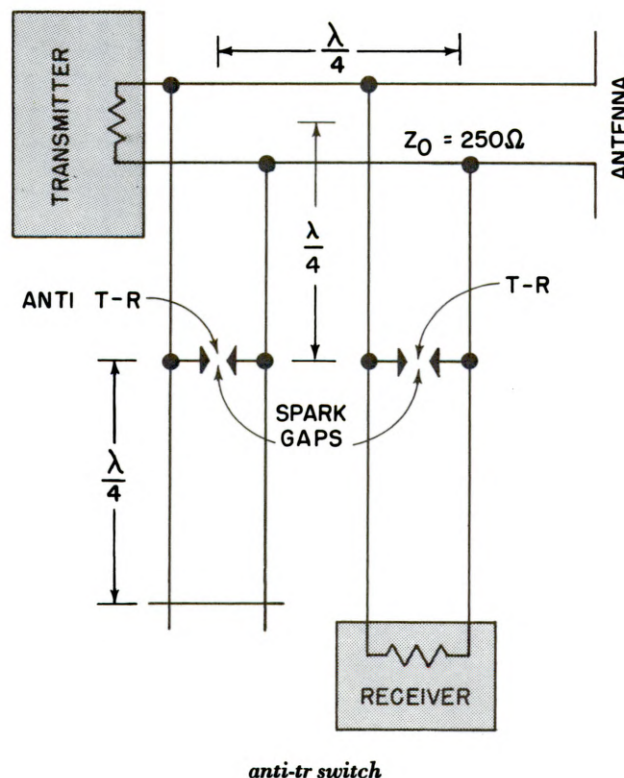
tr and anti-tr devices

In a radar set, the transmit-receive switch is a part of the transmission line structure. Whenever a single antenna is used for both transmitting and receiving, some means must be provided to keep the high power output of the transmitter from entering the receiver circuitry during the transmit portion of the cycle and to provide a path to the receiver for the echo signal during the receive portion of the cycle, while preventing an energy loss to the transmitter. The TR device performs the first of these functions, and the anti-TR device performs the second. As a result of the carrier frequencies employed, no mechanical switching arrangements are available which can effectively operate within micro-second limitations.



The simplest and most basic device that can be substituted for a mechanical switch is a spark gap because it acts as an open circuit (infinite impedance) until sufficient voltage is applied to cause arcing, at which time it has the characteristics of a short circuit (zero impedance).

For proper operation of the device, the impedance at various points along the transmission line must be adjusted to predetermined levels during the transmission of pulses and the reception of echoes. This is accomplished by properly locating the TR device in the branch line to the receiver, and by the use of impedance-changing devices at optimum points along the transmission line.



receivers

The function of a radar receiver is to amplify the comparatively weak echo pulses received from the antenna system, and feed the video output of the receiver to the indicator so that it may produce a visual indication of the desired information. The power contained in the echo pulse that returns to the antenna from distant objects is microscopic when compared with the power of the transmitted pulse. The energy potential that the radar antenna can discern may be a few millionths of a millionth of a watt. The radar receiver must amplify the small, minute voltages several million-fold. After this amplification, the alternating voltage is detected or converted to a d-c pulse. The rectangular outline of such a pulse produces a pip on the range indicator or a spot on the PPI scope. The receiver must accomplish this amplification with the least possible introduction of noise or other disturbances which may be present.

With the exception of some superregenerative IFF receivers, all modern radar receivers are superheterodyne.

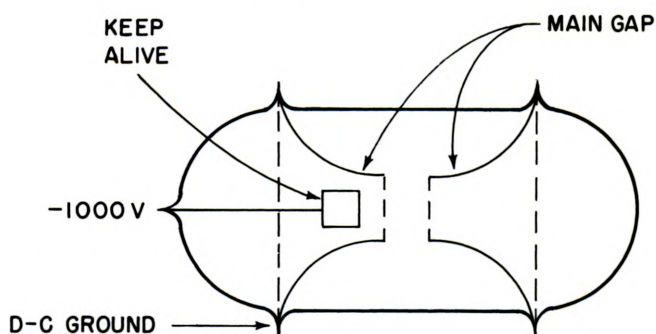
The superheterodyne receiver is used almost universally because it has the advantage of providing a higher overall gain with fewer tubes than other types of receivers. It also has the important inherent advantages of low noise, efficient amplification (pre-tuned circuits), reliability, and freedom from saturations (blocking). The components of a typical receiver are a local oscillator, a mixer, IF amplifiers, a video detector, video amplifiers, and output provisions for built-in or remote indicators. In addition, automatic frequency control (AFC) and automatic gain control (AGC) circuits are used to increase the sensitivity and efficiency of the receiver.

For convenience the transmission lines are drawn in the illustration as parallel wave lines; actually they may be any two conductor type or wave guide type of transmission apparatus.

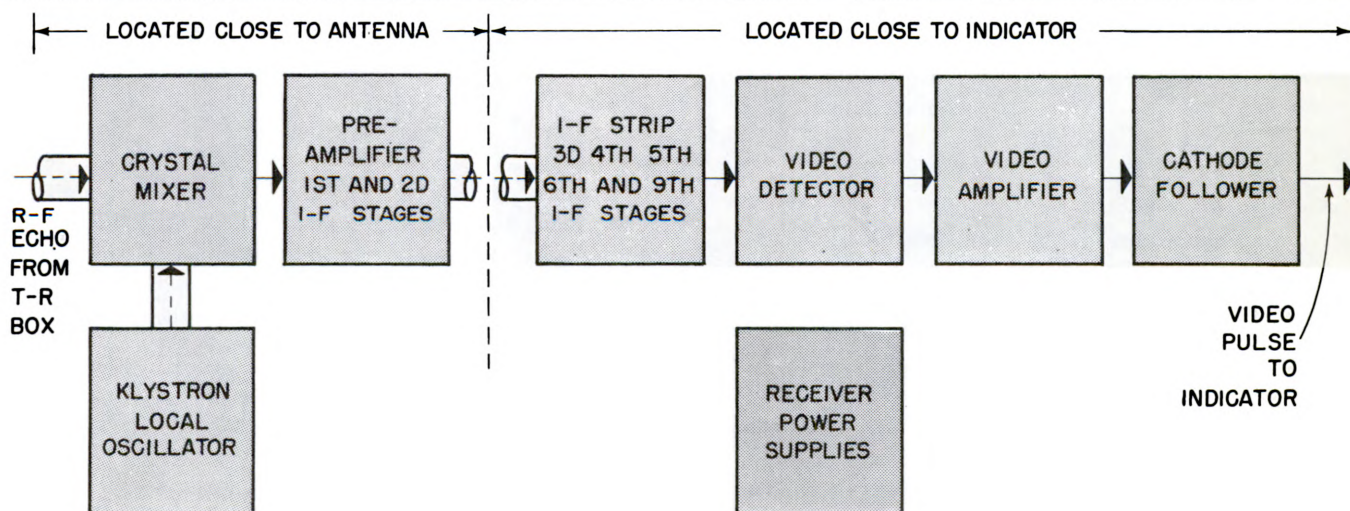
When the transmitted pulse is sent down the RF line toward the antenna, it sees two junction points at which it may subdivide. When it reaches the first junction consisting of the anti-TR switch and associated line, a small portion of its energy is diverted to activate the spark gap. The short circuit impedance of the spark gap is then reflected back to the main line as a point of high impedance, forcing the transmitted pulse to continue down the line toward the antenna. When the traveling pulse reaches the T junction containing the receiver and the TR switch, identical conditions exist and the transmitted pulse continues on to the antenna with very small loss of energy. The anti-TR and TR switches require some of the transmitted power to operate them and although only a small portion of the pulse power is required to perform their functions the switches still are a source of power loss. When the minute energy level of the received echo signal is picked up by the antenna, it is sent back along the same transmission line path as that traveled previously by the transmitted pulse. The received echo signal also sees two paths when it reaches the T junction containing the receiver branch. The anti-TR spark gap represents an open circuit or high impedance path to the traveling wave and very little loss of energy signal to the transmitter occurs. The receiver branch presents an impedance match to the wave guide and therefore receives a maximum power transfer to its input circuits.

The spark gap used in a given system must have a very high resistance until arcing occurs, and then a low resistance during conduction. At the end of the transmitted pulse, the arc must be extinguished as rapidly as possible to allow echo signal flow to reach the receiver.

Enclosing the spark gap in a glass envelope and reducing the atmospheric pressure around the gap has the effect of reducing both the breakdown and running voltages of the gap. The recovery or deionization time of the gap can be reduced by introducing a third electrode mounted near one of the gap terminals. A negative d-c voltage is applied to it to maintain a steady glow discharge near the TR gap. This electrode is known as a keep-alive. During the discharge period electrons from the keep-alive electrode flow to the gap electrode, and because of this action the gap is fired more quickly. The negative voltage of the keep-alive also prevents stray ions from reaching the main gap and producing noises in the receiver. When the radar system is turned off, no keep-alive voltage is available, and it is extremely possible for high-frequency transmitters operating in close proximity to have a portion of the radiated energy detected by the antenna and directed to the receiver section of the local system, providing a possibility of receiver crystal damage. To avoid this possibility a metallic shutter is provided in the receiver wave guide branch. This shutter closes automatically when keep-alive voltage is removed, acting as a protective device for the sensitive receiver.



tr tube with keep-alive



local oscillator

Because of the difficulty in obtaining amplification at high frequencies, substantially all radar receivers employ a superheterodyne circuit on which the signal pulses are connected to pulses of lower frequency, and then amplified. The change of frequency requires the use of a local oscillator to generate a signal that differs from the RF signal by an IF frequency. The two frequencies are beat together in a nonlinear mixer element to produce a difference frequency or IF frequency pulse.

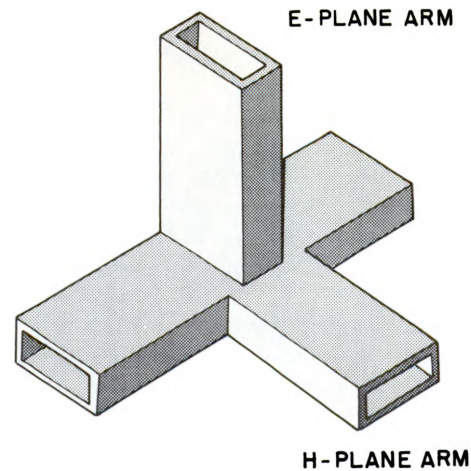
The local oscillator, usually a reflex klystron or light-house triode, operates at a frequency which differs from the RF frequency by 30 mc or 60 mc. When the local oscillator beats against the RF signal an IF output is obtained. The reflex klystron is easily tuned or controlled in frequency by controlling its repeller voltage, and almost all automatic frequency controls take advantage of this fact. Airborne radars usually have a second local oscillator for beacon reception. The beacon local oscillator operates at a frequency different from that of the signal oscillator in order to receive beacon signals. There is no preamplification of the RF carrier and the local oscillator output and the RF echo signal are coupled into the resonant cavity mixer chamber.

mixers

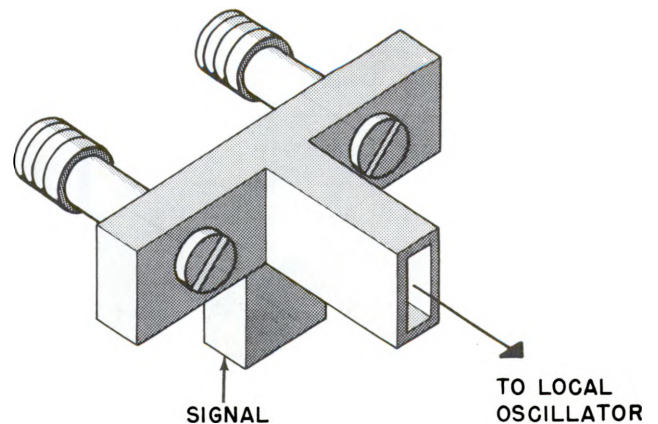
No RF amplifier stage was indicated because available tubes introduce excessive noise when operated above 1000 mc, and most modern radars operate in excess of this carrier frequency. A mixer is a nonlinear device that when fed two signals of varying frequency will produce a beat or difference frequency as its output. The echo pulse is applied directly to the silicon crystal mixer which introduces little noise but does attenuate the signals. The local oscillator output is fed to a second grid of the mixer stage. To minimize the affects of attenuation the mixer is located in the TR box or very close to it. A disadvantage of the direct feed between TR and mixer is that the TR must be extremely effective to prevent damage to the crystal mixer. Circuits that are used to overcome the attenuating characters of the mixer and the inherent noise potential of this stage are balanced mixers.

BALANCED MIXERS

Many airborne radars use a double-crystal balanced mixer. A balanced mixer uses two crystals which are driven in parallel by the local oscillator and in push-pull by the signal. The IF amplifier that follows is designed to respond only to the push-pull output. Noise originating in the local oscillator appears in the same phase at the output of each crystal and is discriminated against. The balanced mixer, which is fed by a magic T, produces an improved signal-to-noise ratio. This is the major advantage of using a balanced mixer.



The magic T junction, as illustrated, consists of a section of wave guide with an H-plane arm attached to the narrow side of the guide and an E-plane arm attached to the broad side of the guide. Because of the orientation of the E and H fields, a wave sent into the H-plane arm excites a wave in the main guide but not in the E-plane arm. Similarly, a wave sent into the E-plane arm excites a wave in the guide but not in the H-plane arm. Therefore the device transmits power to two lines from each of two independent inputs that are not coupled together. By connecting the crystal mixers, as illustrated, the two crystals can be excited by the local oscillator and the signal.



Other advantages of a balanced mixer are: 1) less local oscillator power is required, 2) the radiation of local oscillator signals from the antenna is reduced, 3) the balanced mixer discriminates against image frequencies, 4) the affect of power that leaks through the TR cavity is reduced, and 5) even order harmonics are balanced out at the output. Some of these advantages make the balanced mixer particularly well suited for automatic frequency control (AFC), since they reduce the possibility of locking the AFC circuit at a wrong frequency.

IF amplifiers

An IF amplifier is a high-gain, easily tuned stage of voltage amplification. The IF section of the receiver often includes up to 8 or 9 stages of amplification. The output of the mixer is amplified by the high gain of these stages. The first two or three stages of IF amplification are usually located very close to the mixer-local oscillator combination to prevent attenuation of the IF signal output and act as preamplifier stages. The received signal is of very small amplitude (a few microvolts); therefore the receiver must have a very high gain.

All IF stages are pretuned for maximum response at the IF frequency and for a selected predetermined bandwidth. Regenerative feedback creates a problem when amplifier stages are cascaded. Shielding, power supply decoupling, and stagger tuning are utilized to provide a desired bandpass characteristic.

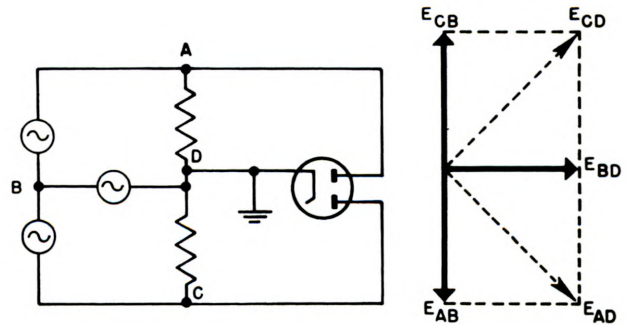
video detectors and amplifiers

The output of the IF stage is converted to video-frequency signals by the detector stage, which is usually a diode. There is normally a slight loss of amplification in this stage. The video amplifier is used to amplify the detected pulse voltage to the required magnitude necessary to operate the indicator. A cathode follower is used to couple the video signal output to the indicator. A characteristic of a cathode follower is its impedance-matching ability. It is necessary to match the output impedance of the video amplifier to the input admittance of the scope, to minimize the noise, distortions and attenuations that would normally occur when a coaxial cable is used to feed signals through to a remotely located indicator.

automatic frequency control

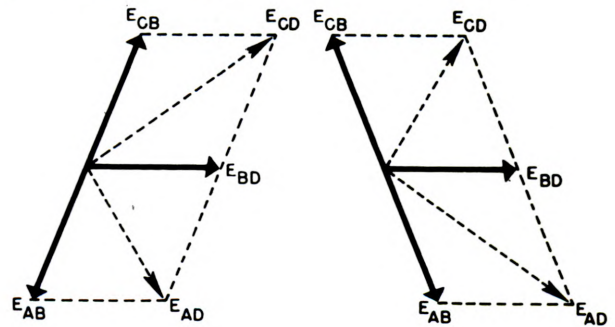
Any frequency drift of the RF signal frequency or of the local oscillator frequency causes the IF frequency to vary proportionately. To compensate for this frequency drift, the IF bandwidth may be increased, but this has the effect of raising the noise level of the set. Other means must be found to correct for frequency drift. Since the frequency stability of both transmitter oscillator and local oscillator are susceptible to frequency shift at ultrahigh frequencies, automatic frequency control circuits are used to stabilize the mixer output.

Automatic frequency control is most easily accomplished by adjusting the repeller voltage of the local oscillator (klystron). The AFC circuit consists of a frequency discriminator, which detects any deviation from the intermediate frequency, and a control circuit, which imposes a corrective change on the local oscillator.



(1) EQUIVALENT R-F CIRCUIT OF A DISCRIMINATOR

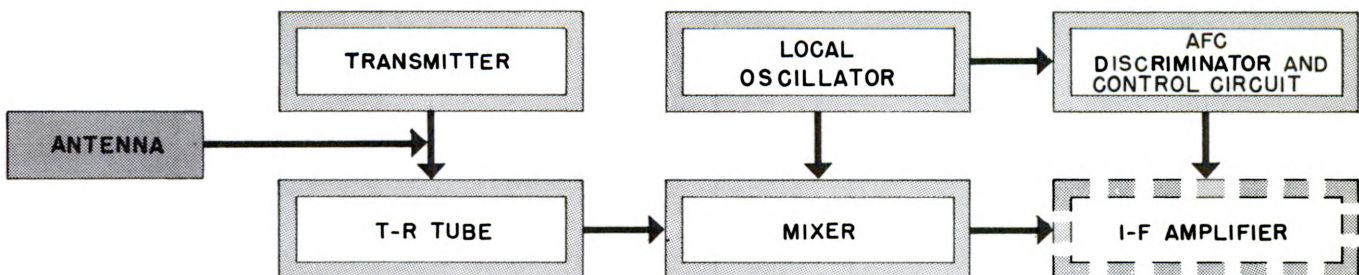
(2) FREQUENCY AT RESONANCE



(3) FREQUENCY BELOW RESONANCE

(4) FREQUENCY ABOVE RESONANCE

AFC circuits are divided into two types: difference frequency (DF) and absolute frequency (AF) systems. A difference frequency system is one in which the difference frequency of two signals is maintained at a constant value irrespective of the absolute stability of either source. An absolute frequency system uses a crystal oscillator and a multiplier to provide an accurate reference signal, which is used to maintain the transmitter or the local oscillator at a predetermined frequency. The difference frequency system is most often used, although absolute frequency AFC circuits are used in radar beacon systems in which beacon transmitters and receiver local oscillators are operated at a single fixed frequency.



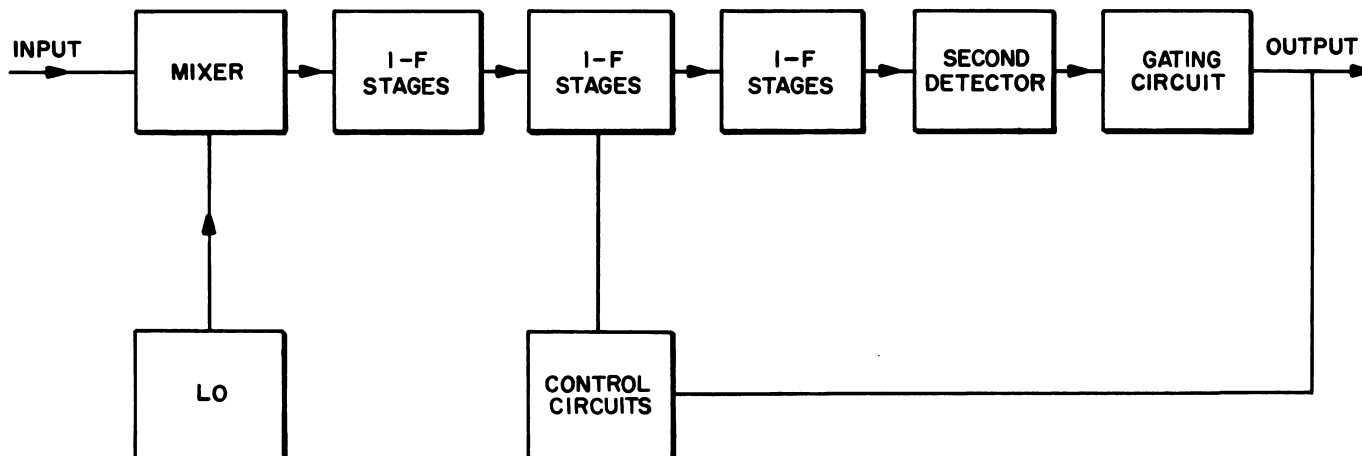
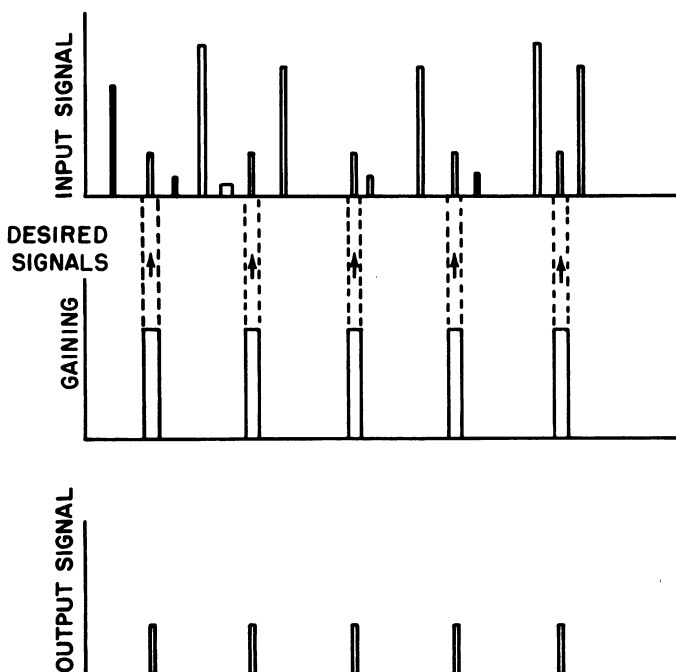
gain control circuits

An automatic gain control circuit has as its main objective the maintenance of the output level at a particular predetermined value. The ideal gain control circuit would automatically maintain a sensitivity such that all echoes from targets of similar size would appear with equal amplitude regardless of the range to the target. This is impossible to accomplish because the power reflected from a target varies inversely as the fourth power of the range. Obviously it is impossible to vary the gain of an amplifier over such wide limits.

A radar receiver is operated at high gain so that small signals of the same order of magnitude as the inherent noise level can be distinguished. Therefore it is desirable to have a gain control circuit that reduces receiver sensitivity when a signal of large amplitude is received. The automatic gain control (AGC) circuit reduces the gain of the receiver whenever a strong signal is received to prevent saturation blocking of the following stages. Detector balance bias (DBB) networks are employed to bias the video detector upon reception of large signals, to insure increased target definition. The correct use of bias circuits prevents overloading and the adding of new frequencies to the pulse waveform, preserving the fidelity of the system. System sensitivity is preserved by the employment of sensitivity-time-control (STC) circuits that reduce the gain potential of the receiver immediately after a pulse is transmitted. This permits the reception of nearby targets without the probability of saturating the IF stage. The problem of automatic gain control of radar receivers is compounded because the signal is a very short pulse. Furthermore it is necessary for the AGC circuit to control the gain of each pulse, since the signals may be large or small and in any order. Such rapidly acting AGC circuits are called instantaneous AGC (IAGC) or instantaneous-acting volume control (IAVC) circuits.

receiver sensitivity

Increasing the sensitivity of the receiver also increases the maximum range of the radar set. The more sensitive the receiver, the longer the maximum range at which an echo signal can be detected. The energy content of the echo pulses that are detected by the receiver is microscopic when compared with the power contained in the transmitted pulses. The ability of the receiver to detect and amplify a minute quantity of energy is a mark of the quality of the receiver and the factor which determines the maximum range of the radar set.



NOISE

Noise in a radar system affects the sensitivity of the system by limiting the maximum range of its operation. Were it not for noise, the maximum range at which a radar system could detect a target would be extended indefinitely. The energy of the echo pulse must be of sufficient volume to prevent the reflected signal from falling below the level of the noise, at which point the radar would lose its ability to display information.

Noise in a radar system can originate at the target, in the environment, in the receiver, or may be inherent in the system itself.

Various types of noise are involved in radar reception. Target noise refers to radiative effects from the target itself, and may be subdivided into two categories: 1) amplitude noise 2) phase or angle noise (glint).

target noise

Target sources such as propellers, exhausts, motor vibration, etc., extraneously modulate (both in amplitude and frequency) the search wave train. They create a degree of noise that is proportional to their energy output. Phase or angle noise (glint) arises because the transmitted wave train slides or creeps over the surface of the reflecting object. The angle of reception of the reflected or echo signal thus differs from the angle of transmission of the original signal. This difference factor may vary from pulse to pulse. Because glint is a function of target size and reflecting surface characteristic its effect normally decreases with an increase in range (as target characteristic at great range is insignificant). Glint is particularly serious at short ranges and small incident angles of detection because variable target reflections under these conditions make the task of separating the target from its image very difficult. The image effect is caused by jitter. Jitter is a relatively small variation of the pulse spacing in a pulse train, creating an image target which is displaced from the actual target by an amount equal to the degree of jitter. Jitter may be random or systematic, depending on its origin.

environmental noise

Environmental noises include all extraneous radiation caused by atmospheric disturbances, both natural and man-made. The frequencies that are generated make this type of noise a factor only at the lowest end of the radar spectrum.

system noise

System noise is usually the result of unwanted vibration, fluctuating thermal agitation, etc. affecting the operating characteristics of transmitting or receiving components. Although these fluctuations must be evaluated, they do not particularly effect the noise level. Even when they do, effects are minute.

input circuitry of receiver

An extremely high percentage of the total noise content of a radar system can be localized in the input circuitry of the receiver. To make possible the detection of targets at great range, the relatively small echo pulse must be amplified to many times its received value.

The sensitivity of the receiver is not limited by the amplification gain that it can be designed for, but by its inherent noise level, which obscures the signal's information content. Noise develops because electronic characteristics of vacuum tubes (shot effect, microphonics, hum, and thermal agitation) are amplified along with the input signal. An added noise source is random pickup which is created by poor component wiring acting as a receiving device. The noise spectrum fills the receiver passband and is considered white noise. The effects of receiver circuitry on noise are difficult to determine accurately because noise is not periodic but irregular and unpredictable. Because of the randomness of the noise, instantaneous noise values have no significance. If all noise sources were considered to be localized in the input circuit (antenna lead into the receiver), then the input receiver noise power could be expressed by

$$R_{NP} = NF \cdot B \cdot K \cdot T$$

where R_{NP} = receiver noise power referred to receiver input

NF = receiver power ratio

B = receiver bandwidth

K = Boltzmann's constant = $1.38 \cdot 10^{-23}$ erg/°K

T = absolute temperature

The minimum signal-to-noise ratio could then be determined as a function of receiver power ratio or reflected power as a function of receiver noise level.

A measure of receiver effectiveness is the signal-to-noise ratio indicated variously as S/N ratio or SNR . The S/N limitation is the reason that an increase in peak power output does not increase the sensitivity of a system. By halving the pulse duration time, peak power is doubled without changing the total energy output. However, to amplify the pulse signals the receiver must be responsive to the frequencies contained in the frequency band. The width of the frequency band required is inversely proportional to the pulse duration time. When the pulse duration time is halved, the frequency band must be doubled. Because noise power is proportional to receiver bandwidth, it is also doubled and therefore there is no change in the S/N .

indicators

An indicator produces a visual indication of the echo pulses in a manner designed to furnish required information content. The parameters desired are range, azimuth or bearing, and elevation of targets.

target range

The range of a target is determined from the measurement of time from transmittal to reception. Increased accuracy in range measurements is obtained by employing calibrated-interval markers (range-markers, notches, and pedestals) superimposed on the sweep. Portions of

the sweep between markers can be expanded to obtain higher resolution in portions of the range sweep time. Receiver sensitivity can be controlled so that target detection can be simplified under adverse conditions.

target bearing

The bearing of a target is a function of antenna beam bearing. Increased accuracy in azimuth measurement is obtained by taking the beam horizontally or by rapid horizontal sector scan, and by indicating signal variations from the lobes on spot or error targets.

RANGE ONLY. The A scope uses a linear sweep applied to the horizontal deflection plates to establish a time base. Since the sweep is linear with time, a calibrated range scale may be superimposed on the tube to indicate range directly. Since the A scope gives information about the shape and intensity of transmitted pulses and echoes, it is widely used in monitoring, testing, and alinement of radars. Variations of the A scope are the J, K, L, M, and R scopes. The J scope uses a circular sweep. The R scope uses an ex-

panded linear sweep in conjunction with a precision timing device that puts markers on the indicator. The M scope has a movable calibrated step or pedestal for ranging. The K and L (pip-matching scopes) are A scopes that give additional bearing or elevation information by comparing (basic comparator) the signals from two antenna beam lobes, either superimposed and offset on type K, or back-to-back, as on type L scopes.

RANGE AND BEARING. A B-type scope plots range against bearing. The sweep voltage is applied through the vertical deflector coils, instead of the horizontal deflector coils, as in the A scopes. The B scope uses a vertical (x) sweep for range and a horizontal (y) sweep synchronized with the antenna rotation. This method of presentation gives range in the x direction and bearing in the y direction. The rectangular sweep of the B scope causes shape distortion, as shown, but gives better resolution of range and bearing. The B type representation of the range and bearing of a small area is called a micro-B display.

The most widely used indicator to give range and bearing is the PPI. It utilizes an intensity-modulated system with a linear radial sweep for range, and a circular sweep (synchronized with antenna bearing) for target bearing. The PPI presentation is sometimes made with the center off to a corner, or side (off-center PPI), with the zero range area expanded (open-center PPI), with a portion of the circular scale expanded (delayed PPI), or with one dimension stretched out (stretched PPI). The off-center PPI with extreme displacement is sometimes called sector PPI display.

RANGE AND ELEVATION. Range and elevation data are presented on E scopes and RHI (range-height indicator) scopes. The E scope has a linear horizontal sweep for range, plus a linear vertical sweep in synchronism with the elevation angle of the antenna. Since the E scope gives the elevation angle of the antenna, lines of equal height are hyperbolas, as shown.

The RHI uses a horizontal sweep for slant range, and a vertical sweep synchronized with the elevation angle of the antenna. However the display of the RHI is expanded on the vertical dimension in such a manner that the lines of constant height are horizontal and equally spaced. The E scope is the vertical analog of the B scope whereas the RHI scope is the vertical analog of the stretched PPI.

BEARING AND ELEVATION. Bearing and elevation data are presented on the C scope. The horizontal sweep is synchronized with the bearing of the scanning antenna, and the vertical sweep is synchronized with the elevation

angle of the scanning antenna. The C scope thus gives the bearing angle and elevation angle of a target. It is used in homing applications in conjunction with a PPI or B display.

ERRORS IN ELEVATION AND BEARING. There are various methods of adding elevation data to range-bearing indicators and of adding bearing data to range-elevation indicators. As has been mentioned for the K and L scopes, the pips from the two lobes of a lobed beam can be matched on an A scope to give either bearing or elevation errors. A B scope can be modified to indicate elevation errors. When a B scope is used with two sweeps, one of which is delayed an amount depending on the elevation angle of the antenna, each target will present two echoes which are in

line coincidence only when the target is in proper (dead ahead) elevation. This modified-B display is known as a double-dot display.







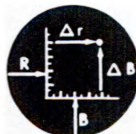

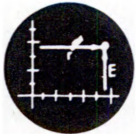
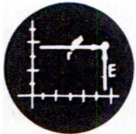






Target alinement (bearing) indications are utilized to measure azimuth deviations from true target position. Variations in echo intensity resulting from a lobed or conically scanned beam are used to move a spot (pip) to indicate bearing deviations. In fire-control radars the spot error indicator is called a train and elevation (T and E) error indicators.

target elevation

The elevation of a target is determined usually from the elevation angle of the antenna. Increased accuracy in elevation measurement is obtained by vertical sector scan or lobing and by indicating signal variations on error or spot scopes. The terms bearing, azimuth, and train all refer to angular distances in the horizontal plane and are often misused. Bearing is the generally used term and may be used to indicate angular distance of a target from a ship's heading (relative bearing), from true north (true bearing), or from any other reference direction. Azimuth is the true bearing of a point, i.e., the bearing from a true north line. Train refers to the relative bearing of gun directors.

display information

Cathode ray tubes are two-dimensional display devices. Thus, only one or two of the desired parameters can be displayed easily on one indicator. The usual indicator is basically the same as a test oscilloscope. The focusing, intensity, and positioning controls are similar. The basic indicators present the following displays: 1) range only (A scope), 2) range and bearing (B and PPI scope), 3) elevation and bearing (C scope), and 4) range and elevation (E and RHI scopes). Different ways of presenting the four basic displays are shown in the accompanying table.

PARAMETER	BASIC TYPE	MODIFICATIONS
RANGE	 A (range in x coordinate)	J (circular sweep for accurate ranging)  M (movable pedestal for accurate ranging)  R (expanded linear sweep for accurate ranging) 
RANGE and BEARING	 B (range in y coordinate, bearing in x coordinate)  PPI (range in r coordinate, bearing in θ coordinate)	Micro-B (x-y representation of range and bearing in a small sectional area)  Off-center or sector display (to expand a region) 
RANGE and ELEVATION	 E (slant range in x coordinate, elevation angle of antenna in y coordinate)  RHI (range in x coordinate, target elevation in y coordinate)	Open-center (to expand small ranges)  Delayed (to expand an interval at a given distance)  Stretched (to increase resolution in one direction) 
BEARING and ELEVATION	 C (bearing in x coordinate, elevation angle of antenna in y coordinate)	
ERRORS in ELEVATION and/or BEARING	 K, L two A traces give errors in elevation or bearing	 MODIFIED B or double dot Spot scopes and Target alignment indicators

CW RADAR SYSTEMS

The basic components of a CW Doppler type radar are a
TRANSMITTER (oscillator),
MODULATOR, ANTENNA,
RECEIVER and INDICATOR.

Basically circuit operation of a CW radar is similar to that of a pulse radar.

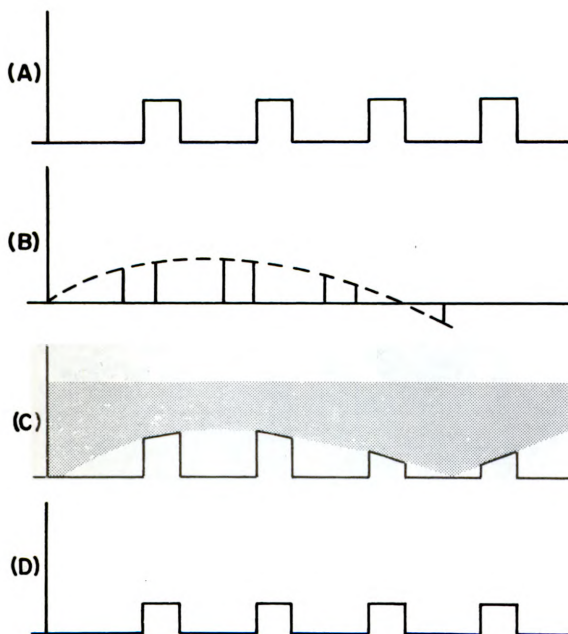
fundamental doppler system

A block diagram of a basic Doppler radar system is shown. The transmitter output frequency of 3000 mc (10 cm) is modulated with the second harmonic output of a 15-mc oscillator. The 3030-mc modulated signal is fed into the crystal mixer stage. The crystal mixer heterodynes the 3030-mc signal with the received signal. The frequency of the received signal is now 3000 mc plus or minus an audio Doppler shift frequency caused by a motion of the target. The resultant frequency is fed to the IF amplifier. The frequency modulator of the transmitter assures a constant intermediate frequency

and avoids tuning problems. The IF amplifier stage, a predetermined high-voltage gain stage, is tuned to a narrow IF bandwidth. The output of this stage is fed to the second detector. The second detector stage utilizes a 30-mc crystal oscillator to beat against the output of the IF amplifier, producing the audio Doppler-shift frequency. The audio frequency is then amplified in the audio amplifier and applied to the indicator. The system as described does not indicate the magnitude of the radial velocity and is a basic simple Doppler Radar system used to indicate changes in velocity.

combination systems

The previously described CW Doppler radar can distinguish between fixed and moving targets and measure velocity, but cannot measure range. Pulse radars can measure range, but they cannot reject clutter echoes from fixed targets. A combination of the two can be used to measure the range and the velocity of a moving target.

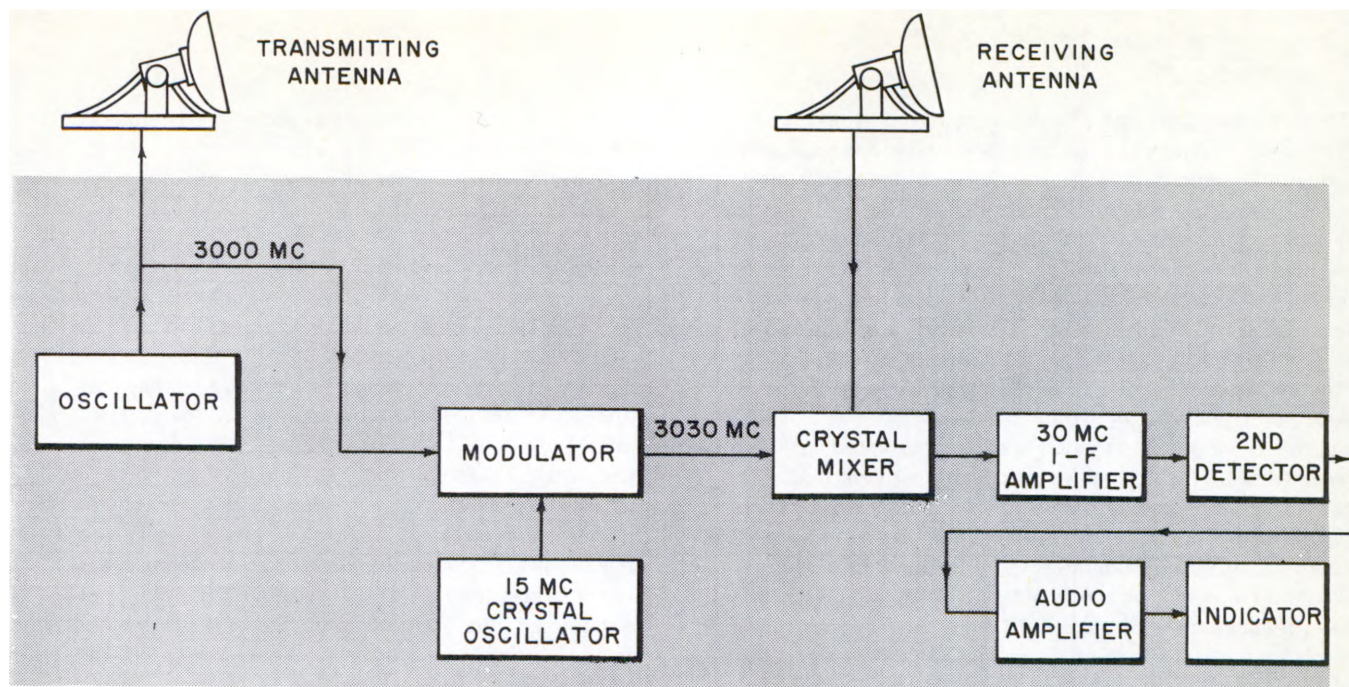


pulse-modulated doppler system

The block diagram of a pulse-modulated Doppler system is shown. Since pulses are used, only one antenna is needed. The square-wave generator, modulator, and TR and ATR tubes control the keying cycle, as in any pulse radar.

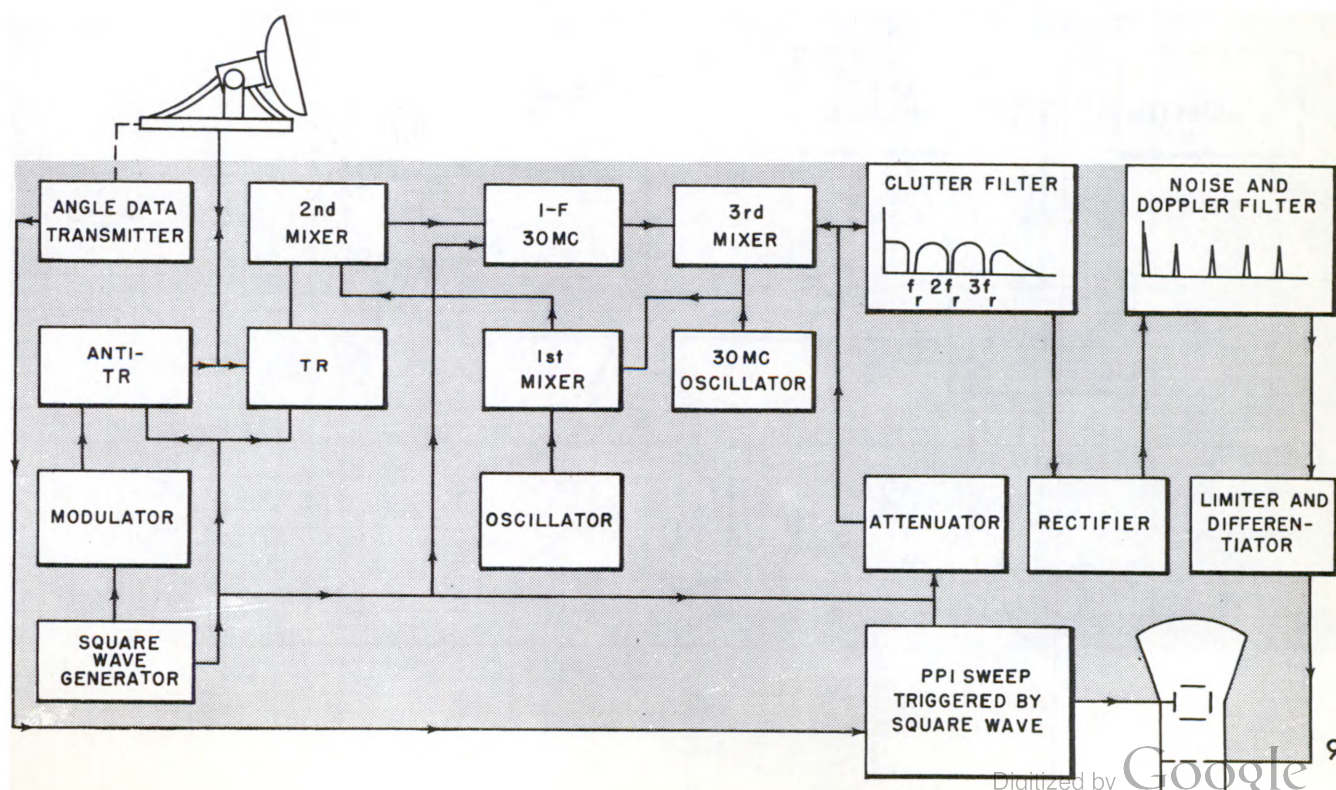
The video pulse-echo waveforms from a moving target are shown. The Doppler frequency appears as the envelope frequency of the waveform. The change in phase from pulse to pulse is caused by the motion of the target. The distance traveled by the target between pulses is vT , where v = velocity of target and T = time between pulses. Therefore, each pulse travels a distance of $2vT$ (or $2vT/\lambda$ wavelengths) less than the preceding pulse. This is equivalent to a phase change of $2(2vT/\lambda)$ between pulses, or to the Doppler frequency.

CURVES SHOWING VIDEO SIGNALS FROM STATIONARY AND MOVING TARGETS, AND RECOVERY OF RANGE FROM MOVING-TARGET ECHOES



Such a combination radar set may be designed either as a pulse-modulated Doppler radar or as a moving target indicator (MTI) radar. If the pulse length is more than 10 percent of the wavelength, the radar set is classified as a CW pulse-modulated Doppler system; if the pulse length is less than 10 percent of the wavelength, the set is classified as an MTI system. The

shorter pulse length of MTI radars gives them greater resolution and permits them to handle more targets simultaneously. However, the longer pulse length of pulse-modulated Doppler radar permits more efficient recovery of Doppler frequencies and better rejection of clutter echoes from stationary targets.



The echo waveform can be explained also in terms of frequency components. A pulsed carrier consists of a carrier frequency plus the components due to the pulse waveform. A rectangular pulse repeated at a frequency f_r has components of $f_r, 2f_r, 3f_r, \dots$ and so forth, that is, nf_r , where $n = \text{any integer}$. If the pulsed carrier is heterodyned to zero frequency, the carrier frequency is translated to zero frequency. However, any other components do not disappear, and the output consists of the video components, nf_r , plus whatever Doppler is imposed on the signal by target motion.

Echoes from stationary targets, as illustrated, have components of nf_r after the carrier has been removed. The echoes from a moving target have a Doppler-frequency (f_d) added to the square-wave frequencies (nf_r). Just as any amplitude-modulated wave is the vector sum of a carrier and one or more sidebands, the sum of nf_r and f_d is a succession of pulses (nf_r) which are amplitude modulated at the Doppler frequency (f_d). The rejection of clutter echoes and the recovery of range information from the Doppler frequency is based on the fact that the echoes from moving targets have frequency components $nf_r + f_d$, whereas echoes from fixed targets have nf_r components.

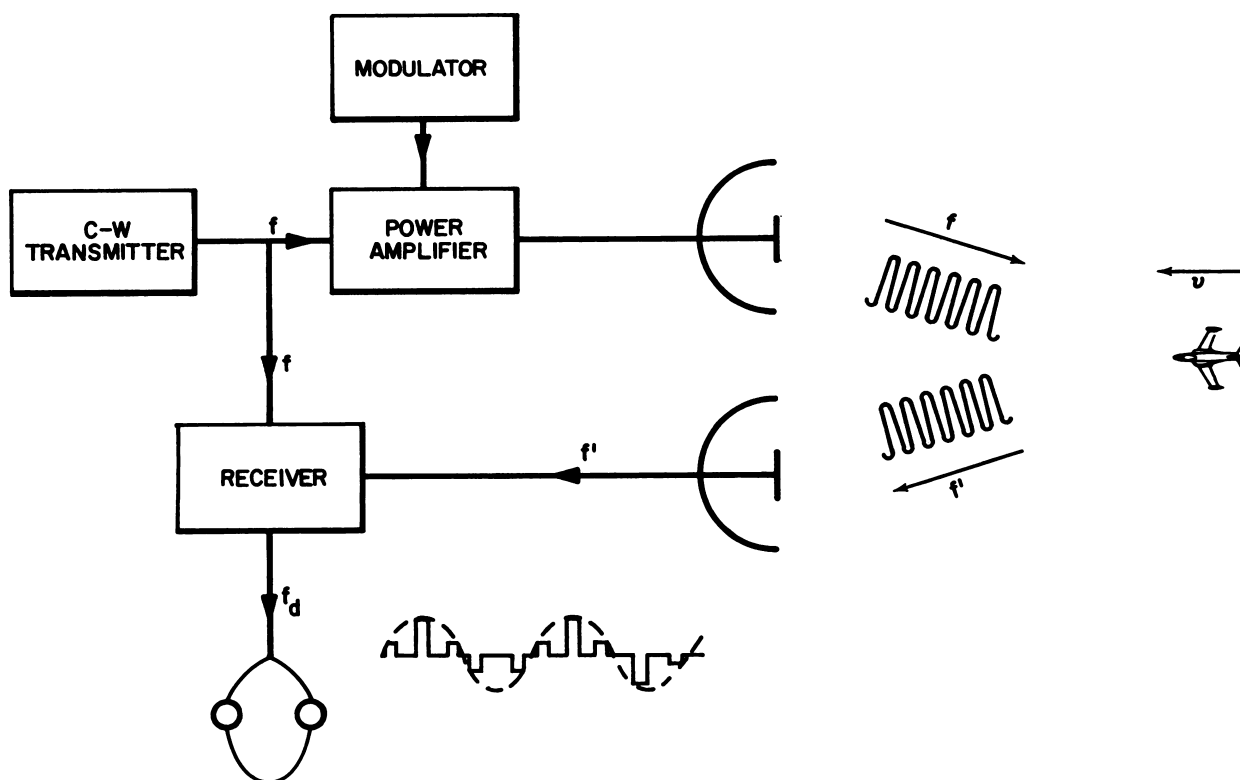
Each echo is heterodyned first to 30 mc, and then to zero frequency, as in the simple Doppler system. The envelope of the output of mixer 3 thus contains only frequencies of nf_r (from the pulses) and f_d (from Doppler effect), so that the echoes from stationary targets are as shown in curve a, and those from moving targets as in curve b.

The signals from stationary objects are rejected in the clutter filter which follows mixer 3. The clutter filter is a series of narrow band cut-off or infinite attenuation filters that reject frequencies of $f_r, 2f_r, 3f_r, \dots$ and so forth. The output of the clutter filter thus consists only of Doppler-frequency pulses, that is, pulses whose amplitude changes from pulse to pulse, as in curve b. Note that the Doppler pulses and the stationary target pulses have slightly different shapes at the tops. The clutter filter can be thought of as recognizing the difference and rejecting the echo from the stationary target.

Next, Doppler-frequency pulses are rectified to give the form shown in curve c; they are then filtered and limited to give the form shown in curve d. The rectification reintroduces the repetition-frequency components (nf_r), and the final waveform has range information in the form of distances between leading edges. The video signal is presented on a PPI indicator to measure the range of the moving target.

Although pulse-modulated Doppler has good rejection of ground clutter, the emphasis in development for naval applications is on MTI. This is due to the better resolution of MTI radar.

Although MTI radars are pulse radars, the principles of MTI and pulse-modulated Doppler radars are so similar that the basic principles of MTI will be discussed at this time.



MTI radars

If the echoes from fixed and moving targets, as illustrated, are presented on an A scope, they appear as shown. The echoes from a moving target vary in amplitude and give the signal a fluttering appearance; the echoes from a stationary target have steady amplitude. The pulse-modulated Doppler system uses filters to reject the pulses of constant amplitude, whereas MTI systems use pulse-to-pulse cancellation of echoes. The first four curves illustrated are the four individual traces shown superimposed on the A scope. If the signals of sweep 2 are delayed for an interval equal to the repetition period and then subtracted from the signals of sweep 1, all signals of constant amplitude are canceled and only variations between successive pulses remain. The last three traces B, C, and D show the canceled signals.

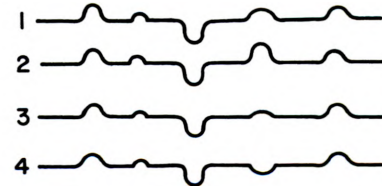
One method of delaying the pulses is to use a supersonic delay line. The subtraction process can be carried out continuously by the arrangement shown. The video signals from the receiver are sent through two channels, one of which contains the delay line. The undelayed echoes and the delayed echoes are then compared in a subtraction circuit, after which the difference signals are presented on an A or PPI indicator. The delay line and subtraction circuits perform the function of the clutter filter and range-restoring elements in the pulse-modulated Doppler system; that is, pulses of constant amplitude are rejected and Doppler pulses of variable amplitude are accepted.

One type of supersonic delay line consists of two electromechanical (piezoelectric) transducers separated by a medium in which the pulse travels at reduced speed. The video signal is introduced between the mercury in the end-cell and that in the delay tube. The voltage across the piezoelectric crystal deforms the crystal. The vibrations are communicated to the mercury in the delay line as a supersonic wave, which travels in mercury at the velocity of 1,700 feet per second. The length of delay line determines the amount of delay. At the output end, the wave causes the crystal to vibrate and produce an alternating voltage. The output voltage is taken from another end cell. The end cells are used as electrodes to introduce and recover the video signals and also to terminate the delay line with the proper impedance.

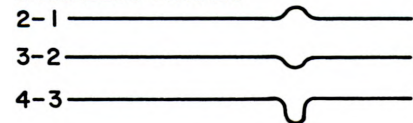
The response of the crystals does not permit undistorted transmission of the video pulses at the input and output. An undamped crystal has too high a Q and cannot transmit microsecond pulses. A critically damped crystal has a Q near unity, with maximum response at the resonant frequency of the crystal and zero response at zero frequency. Whether the crystal is damped or not, a video pulse in the shape of a square wave would be distorted. To avoid distortion, the pulses are used to modulate a carrier, the frequency of which is the resonant frequency of the crystal.



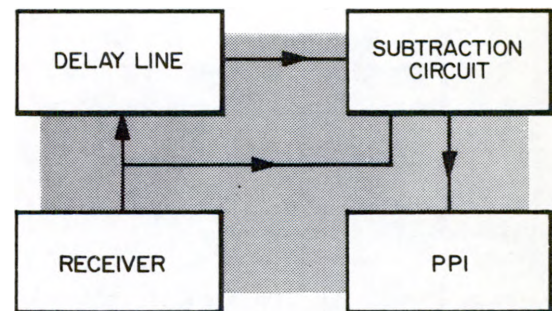
fixed and moving targets



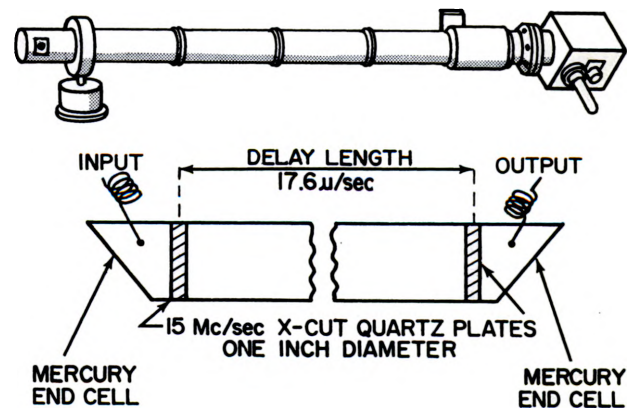
CANCELLED SIGNALS



cancellation of echoes

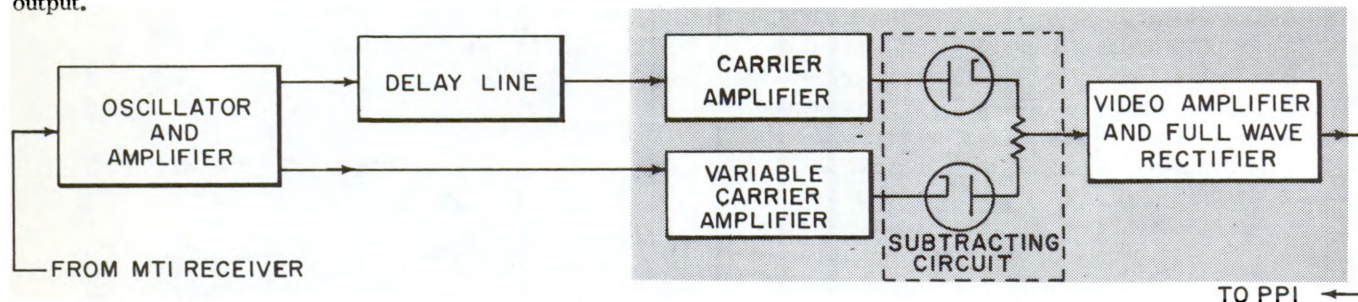


cancelling signals by subtraction



supersonic delay line

The video signals from the MTI receiver amplitude modulate a 15-mc oscillator. The amplitude-modulated carrier is sent through the delay line and then subtracted from the undelayed carrier in the subtraction circuit. The subtraction circuit consists merely of two diodes arranged to give opposite polarities at the output, so that signals of equal amplitude produce zero output.



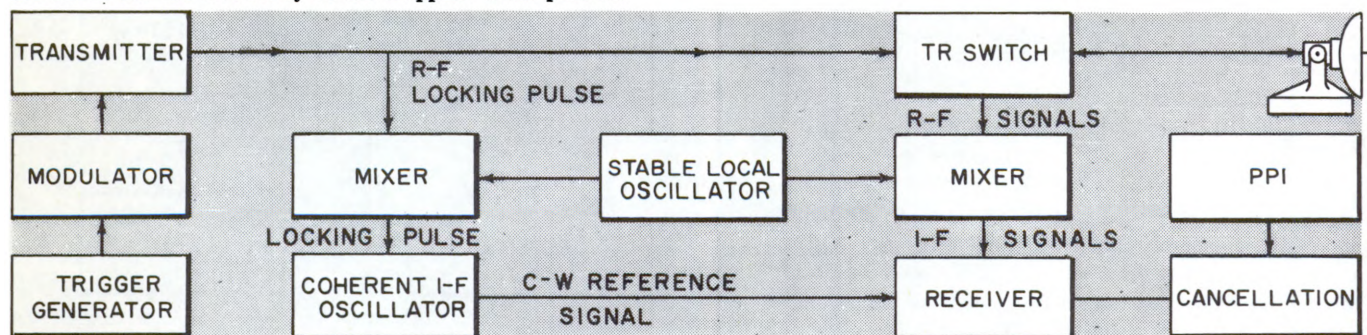
A complete MTI system is shown. As in the simple Doppler system, the echoes are translated to zero frequency by heterodyning in the receiver at the intermediate frequency. The CW reference signal that is used in the heterodyning process is obtained from the coherent IF oscillator.

When pulsed, the transmitter starts with random phase from pulse to pulse. The phase of the reference oscillator must be matched, or locked to that of the transmitter at each instant of pulse transmission, since the phase change beating oscillator and returning echo is the basis for the recovery of the Doppler-shift pulses.

In addition to the supersonic delay line, another device for delaying the echo signals in an MTI system is a storage tube, which is similar to the iconoscope used in television. However, mercury or quartz delay lines are used more often in MTI systems that are now being tested.

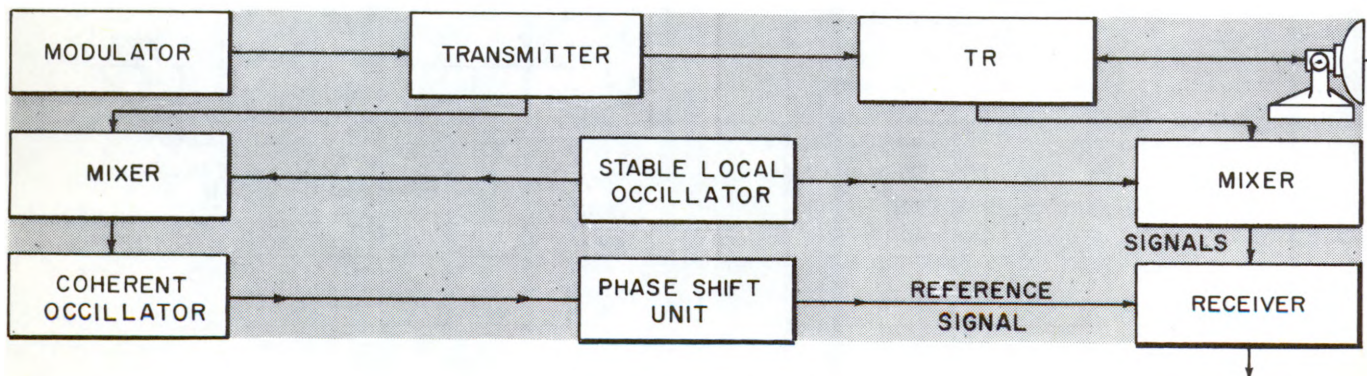
An RF locking pulse from the transmitter is used to phase the oscillator. The locking pulse is said to make the oscillator inherent with the transmitter, and the oscillator is called a coherent oscillator or coho.

Various MTI systems are under development for naval purposes. The systems differ in the methods employed for interlocking the transmitter and coho at RF or IF frequencies, and in the types and location of the delay and subtraction circuits employed.



Moving-target indication circuits can be added to a radar aboard planes or ships by compensating for the velocity of the station. The compensation can be done by a phase-shift unit. The phase-shift unit shown changes the phase of the reference signal at the same rate that the phase of echoes from fixed targets is being changed by the motion of the station. For air-

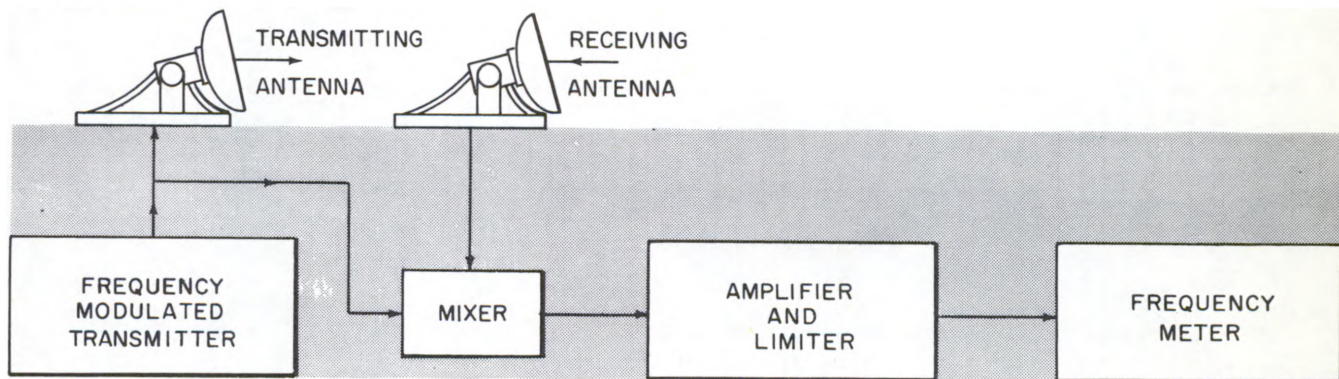
borne sets, an added difficulty arises because the radial velocity between the plane and fixed targets depends on the depression angle, and the phase-shift unit cannot be used for clutter below the plane at large depression angles. This necessitates the use of different systems, such as noncoherent-oscillator types.



FREQUENCY-MODULATED RADAR

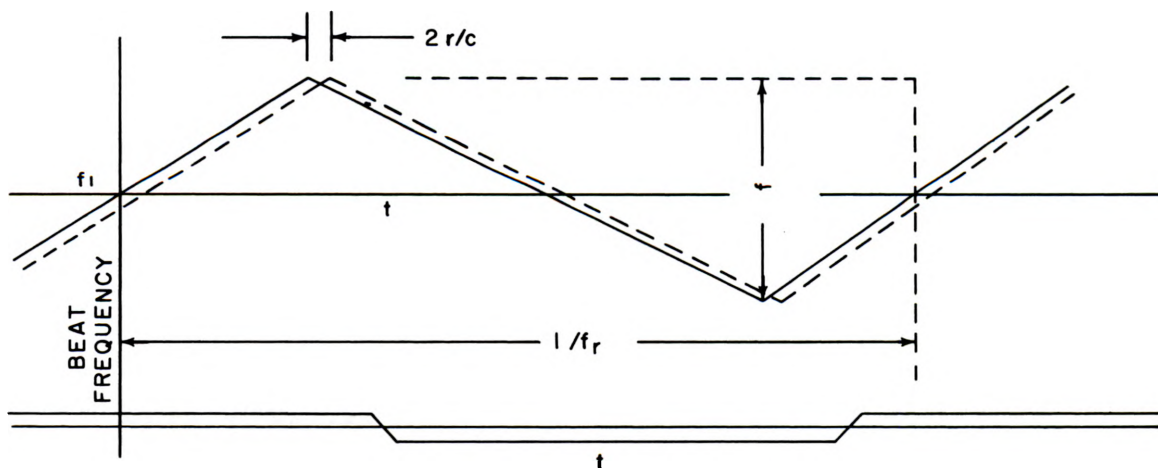
Frequency-modulated (FM) radar systems transmit a continuous frequency-modulated signal. A target returns a delayed signal, which at any one instant differs in frequency from that of the transmitter by the amount of frequency change imposed on the carrier in the delay interval. The difference frequency depends on the delay interval and, therefore, on the range.

Separate antennas are used for transmission and reception so that no time delay is introduced for switching a dual-purpose antenna. This permits measurement of ranges as short as a few feet. The use of a CW signal means that the bandwidth, the rate of information collection, and the resolution are much smaller than for pulse radar. Therefore, FM radar is useful primarily where one important target is illuminated and where speed of indication is not critical. The principal application of FM radar in the Navy is in the radio altimeter, with reflection occurring from the surface of the ground or ocean.



Where a single reflecting object is responsible for the echo, the beat frequency is recovered and used in remarkably simple circuits to indicate altitude, to control aircraft in level flight, and to control missile release in bombing. Doppler frequencies caused by target motion are pronounced in the FM system, so that measurement of velocity as well as recognition of moving targets in the presence of clutter is easily performed.

The block diagram of an FM system for measuring range is illustrated. The transmitter emits a wave the frequency of which varies with time, as shown by the solid line. The mixer receives two signals, one from the transmitter, and one from the receiving antenna. The latter signal lags the former by the interval required for the wave to reach the target and return. The output of the mixer is the difference, or beat frequency between the two signals. The beat frequency is amplified, limited, and measured by a cycle-counting device that is calibrated in terms of range.

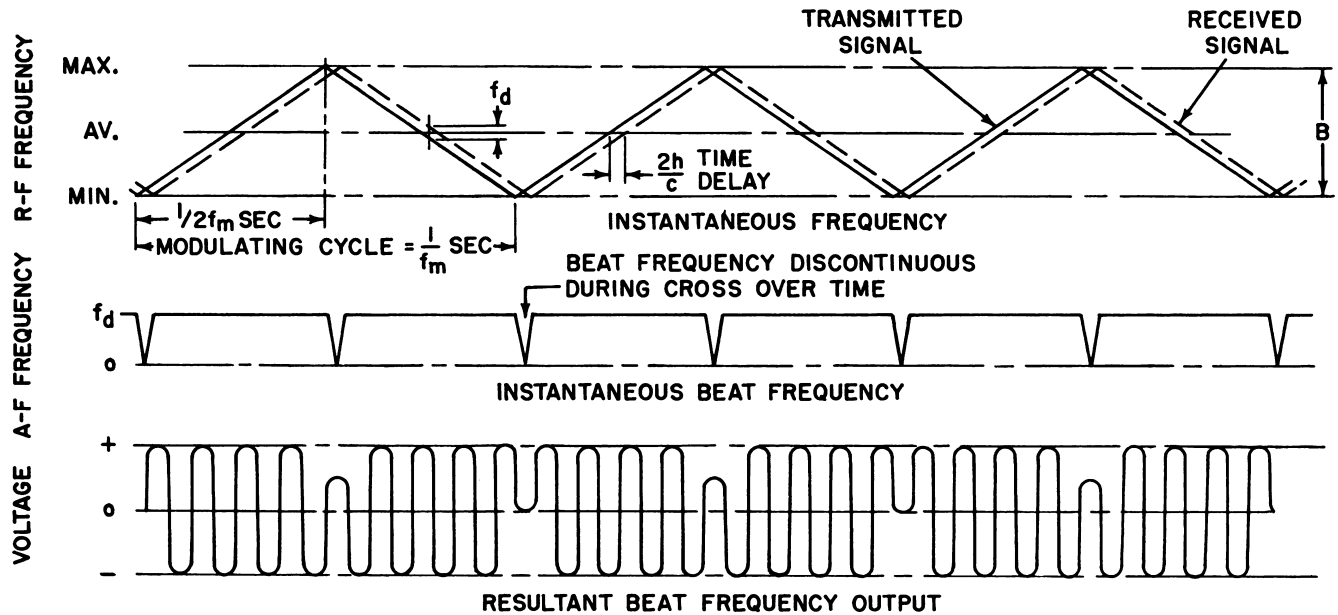


As shown, the echo signal changes alternately from a value above the transmitter frequency to one below the transmitter frequency. However, the audio frequency output of the mixer is the average value of the beat frequency (f_d). The beat frequency depends on the range (h), on the number of times per second that the transmitter is modulated (f_m), on the frequency deviation

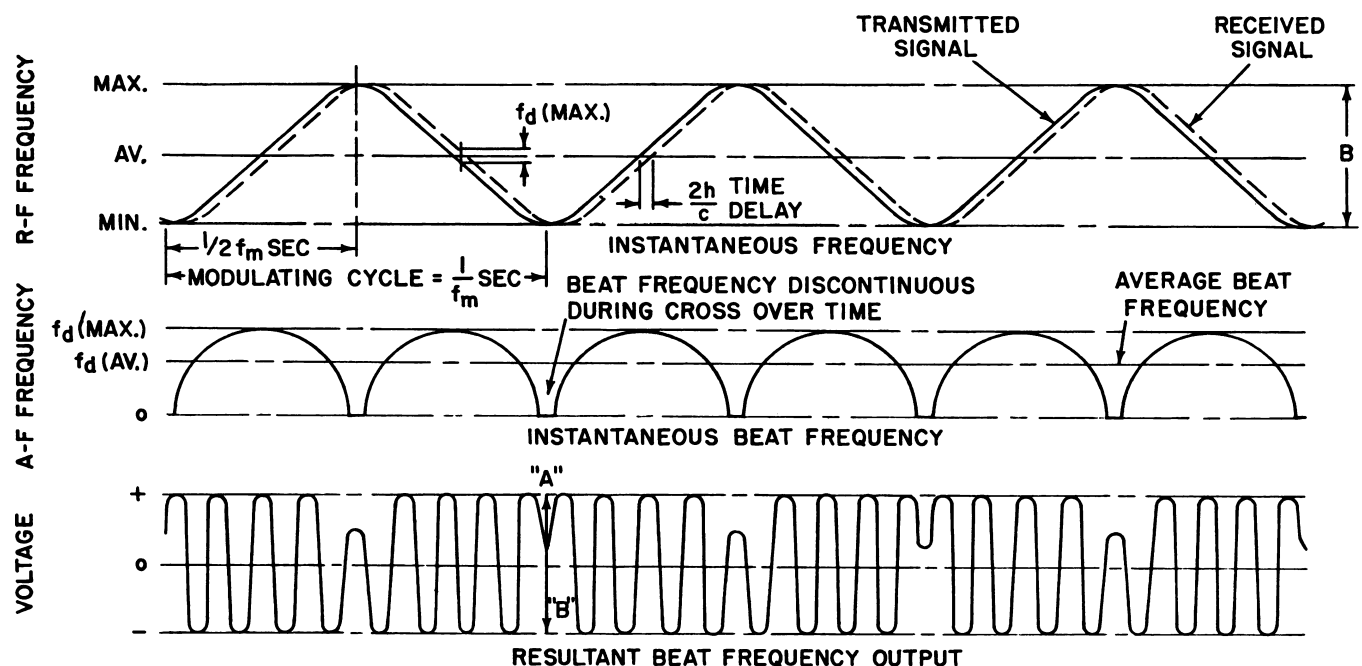
imposed on the carrier (ΔF), and on the velocity of the wave (c), as follows:

$$f_d = 2f_m (\Delta F) \frac{2h}{c},$$

The resultant beat-frequency output for symmetrical sawtooth frequency modulation of the transmitter is shown.



If the transmitter is modulated sinusoidally instead of by a symmetrical sawtooth, the instantaneous beat frequency and the resultant output are as shown. Fortunately, the average value of the beat-frequency output is exactly equal to the constant beat-frequency difference for sawtooth modulation. Sinusoidal modulation is accomplished easily and is used in radio altimeters.



RADAR COUNTERMEASURES

Radar countermeasures are designed to deny the enemy the use of his radar, to cause the enemy to be deceived by his own radar, and to counter the enemy's attempts to jam or deceive friendly radar. The necessary information regarding the characteristics of enemy radar are obtained by the use of intercept receivers, panoramic adapters, pulse analyzers, and direction finders. Intercept or search receivers are designed to scan the possible frequency bands that the enemy may transmit on, and to detect his presence. Panoramic adapters and pulse analyzers are designed to analyze the video output of the search receivers, and to provide means for measuring the pulse width, repetition frequency, and other pertinent information that may be contained in the intercepted pulse. Direction finders are utilized to locate the enemy radar.

After the desired information is computed, tactical considerations determine whether the enemy radar is to be evaded by using blind approaches, deceived by mechanical reflectors such as chaff, window, and decoys, or jammed by employment of an electronic jammer. Usually a combination of tactics is used.

nonelectronic devices

window

Window is the term applied to the strips of aluminum cut to such a length that they are resonant at the frequency of the radar to be jammed. When dropped from planes, a packet of such strips returns echoes similar to those from aircraft. The length of the strips determines the frequency response of the countermeasure device. A further limitation of this device is that the strips must drop or move in linear or slightly curved motion to confuse the enemy radar effectively. If the strips are made too small or too light, they will drift in space and their purpose will be easily detected. Such strips, because of the weight and length limitations, cannot be made to reflect signals transmitted at frequencies higher than 500 megacycles. With the advent of microwave radar, the use of window as a countermeasure device has declined.

rope and decoys

For employment against long wave search radar, long ribbons or ropes of reflecting foil are more efficient than windows. Although untuned, the foil twists and turns and presents many aspects to the radar, some of which appear as strong target echoes. Corner reflectors and radar decoys (balloon-borne streamers) are other mechanical devices which return a strong false echo and appear as large targets on enemy scopes. Their use as decoys for deception is obvious.

electronic devices

An electronic decoy is a type of transponder that returns a fairly large signal each time it is triggered by an enemy pulse, confusing the enemy radar, and making him think a target exists at the location of the transponder.

transponders

The most familiar type of transponder is the basic transmitter which is located in a radar beacon installation (racon). This type of radar beacon is employed for direction finding, triangulation, and for identification purposes. There are many dissimilarities between IFF transponders and racons including coding methods, frequency bandwidth, and circuits. Normally IFF transponders operate over a relatively narrow frequency range, while racon transponders operate over a wide frequency range. Regardless of circuitry the basic objective of the decoy transponder is to emit a false echo pulse to deceive the enemy radar.

IFF principle

Identification friend or foe (IFF) radar is a necessary adjunct to a radar system utilized to identify a target. When a radar operator detects a target on his indicator he can interrogate the target by means of interrogating pulses from his IFF transmitter. When the IFF interrogation pulse reaches a friendly plane, a transponder is triggered. The transponder emits a characteristic coded signal that is returned to the receiver of the IFF system.

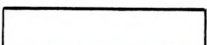
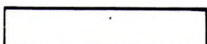
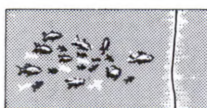
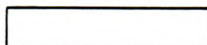
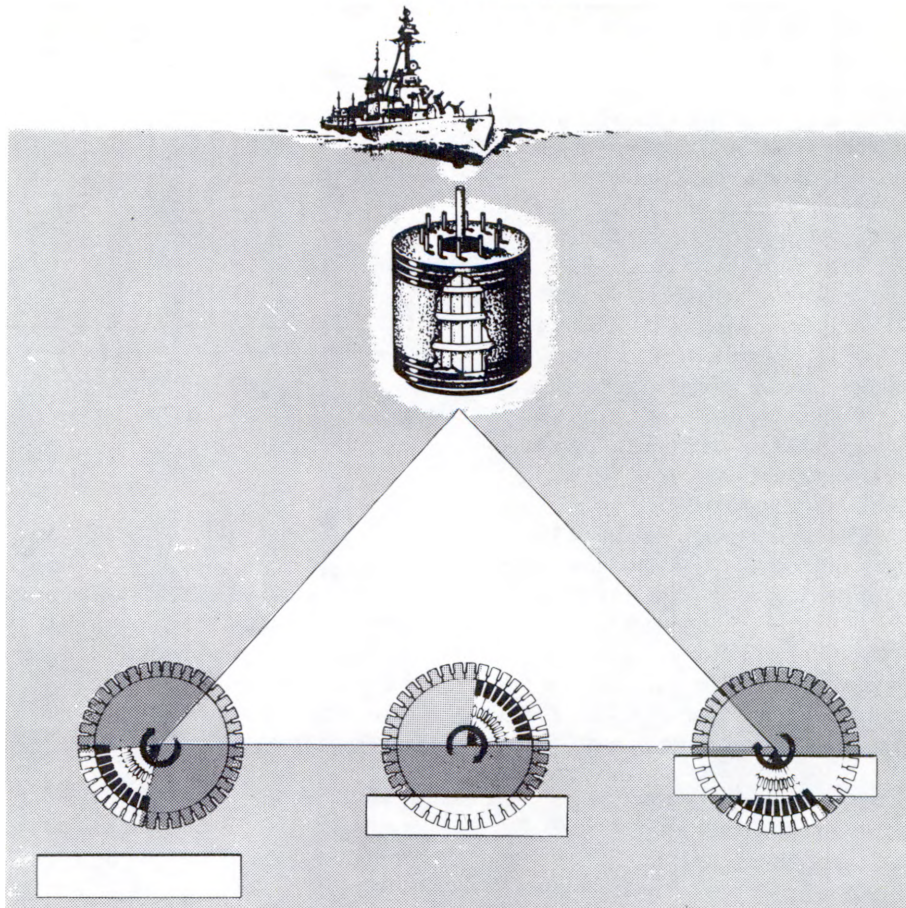
tactical considerations

Radar countermeasure equipment (RCM) is used tactically in many ways other than in locating and jamming an enemy radar installation. Various other capabilities of countermeasure equipment are listed:

- 1). Detecting and ranging of enemy radars
- 2). Predicting enemy action through observation of his radar signal
- 3). Plotting the antenna field pattern of enemy installation
- 4). Determining the blind approaches to a target area
- 5). Obtaining navigational fixes from known enemy or friendly radar installations
- 6). Setting jammers on the frequency of enemy radar signals
- 7). Deceiving the enemy regarding our true activities
- 8). Identifying friendly craft by identifying signals
- 9). Providing direction finding bearings on homing signals for friendly aircraft or missiles.

Many of the tactical uses originate from the fact that a radar signal can be picked up by an intercept receiver at distances far beyond the effective range of the radar. Intercept missiles can then be homed on enemy radar. Tactical considerations are important in combating the countermeasures of the enemy. The effects of jamming can be reduced considerably by proper use of radar controls and antijamming (AJ) filter circuits. Evading enemy radar is an important tactical consideration in which RCM intercept equipment is employed. The pulse width of the enemy radar determines the minimum range and resolution. The frequency and power of the radar and the terrain features determine shadow areas that may be used to evade the radar. The shadows caused by the earth's curvature obscure low-flying aircraft at long ranges. Electronic and mechanical jamming should be avoided if evasion is possible. The increased efficiency of counter-countermeasures against enemy countermeasures will prove to be one of the most important aspects of future military development.

SONAR



The term sonar is derived from the words

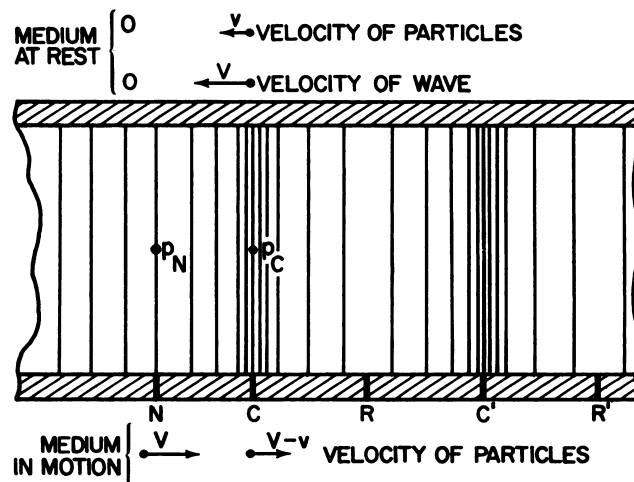
Sound..... Navigation . . . Ranging

This term is applied to systems in which underwater acoustic energy is used for observation, detection, or communication. The term is also applied to the principles and practices employed in the design of such systems. Sonar may thus be seen as that branch of applied acoustics which is identified by the use of water as the propagating medium. The development of the submarine and consequently of antisubmarine warfare (ASW) has created the necessity for devising more expeditious methods of underwater detection. Submarines can be detected by purely visual means, or by means of measurement of magnetic anomalies, by pressure-sensing devices, by detection of electromagnetic radiation or of electromagnetic action. The use of underwater acoustics in submarine detection, however, appears the best means thus far devised and promises to remain the most fruitful for years to come.

PHYSICS OF UNDERWATER SOUND

propagation

The term sound is used in two senses: subjectively it signifies the auditory sensation experienced by a receiver diaphragm (ear), and objectively it signifies the vibratory motion which gives rise to sound wave motion. It is used in the latter sense in sonar. Vibratory action is the essence of sound wave propagation. All wave motion can be classified as either longitudinal or transverse in character. A longitudinal wave is one in which the vibrating particles move forward and backward parallel to the direction in which the wave is propagated. A transverse wave is one in which the particles vibrate at right angles to the direction of propagation. Sound waves are longitudinal in character, and their progress through a medium involves two distinct motions. The wave itself in a homogeneous medium moves forward at constant speed, which means that the configuration advances equal rates of movement in equal periods of time. At the same time the particles of the medium that conveys the wave vibrate to and fro in harmonic fashion. Their locations at successive moments depend upon the period, amplitude, and phase of the vibration. The period (T) of a vibrating particle is the time in which a wave completes one vibration, and the frequency is the number of vibrations completed per second. The amplitude of the vibration is the maximum displacement from the undisturbed position and is a measure of wave intensity.



Waves transmit energy in the direction of propagation; hence, a continuous transfer of energy takes place in the direction of wave travel. If a wave diverges as it advances, the amplitude of the vibration diminishes as the wave progresses, since the energy is distributed over a larger surface area. This process is called wave energy diffusion.

intensity of sound waves

The intensity of sound is the time rate of transfer of vibratory energy per unit of sectional area of the sound wave or the rate at which sound energy flows through a unit area perpendicular to the direction of the wave. Both kinetic and potential energy are present in a sound wave, and the average kinetic energy equals the average potential energy. Assuming the sound wave has simple harmonic motion, and that the unit area considered is thin enough to assure that all the particles in it have equal displacement, then if x is the thickness of the layer, and d is the density of the medium, the mass of this volume, m , equals xd . The maximum kinetic energy, E , of the area under consideration then is

$$E = \frac{1}{2} mv^2 = \frac{xd}{2} (2\pi fr)^2$$

where f = frequency of the vibrating particles
 r = maximum displacement
 v = maximum velocity

The energy of the layer is in kinetic form when the particles are in equilibrium position and in potential form when the particles have maximum displacements. In consequence, the total energy per unit volume is $E = 2\pi^2 f^2 r^2 d$ and may be termed the total energy density

of the wave. When metric units are employed, E is in ergs per cubic centimeter. If the velocity of the wave propagation is in centimeters per second, the time rate of transmission of energy per unit area will be the velocity times the energy density of the sound wave. Thus the intensity of the sound in ergs per second per square centimeter is

$$I_r = 2\pi^2 r^2 f^2 v^2 d \frac{\text{ergs}}{\text{cm}^2 \text{sec.}}$$

From this relationship it is seen that the intensity of sound is proportional to 1) the square of the amplitude, 2) the density of the medium, 3) the velocity of propagation, and 4) the square of the frequency of vibration. The basic equation for determining energy density demonstrates a vital point of sound transmission characteristics. At any distance from a point source, the energy density will vary as the square of the distance. Also, as a sound wave moves through a medium, variations in pressure occur at all points in the medium. The greater the pressure variations, the more intense is the sound wave, and sound intensity is proportional to the square of the pressure variations regardless of the frequency.

refraction

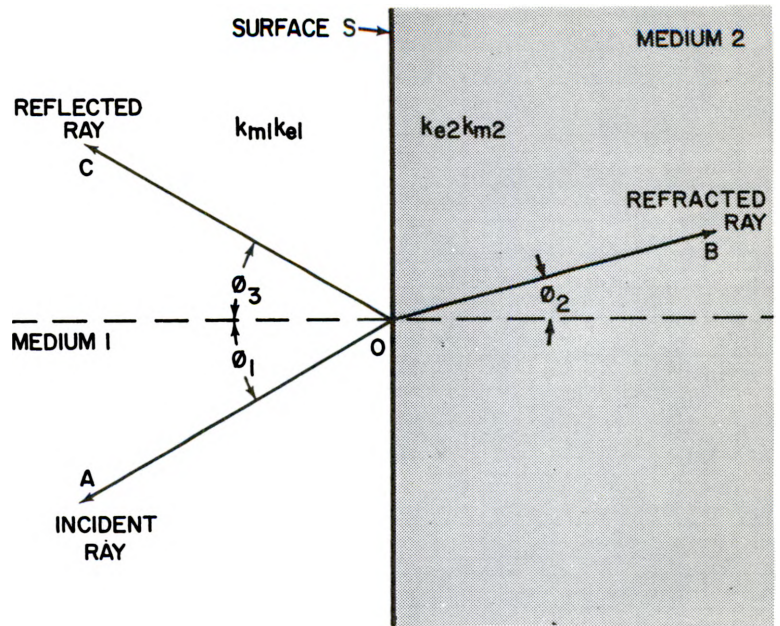
A wave which enters another medium or another layer of the same medium having different characteristics will undergo an abrupt change in direction and in velocity. In the illustration given, an incident ray, A, is shown impinging on a surface, S, dividing two media of differing characteristics. The wave velocity, v_1 , in medium 1 and the wave velocity, v_2 , in medium 2 are related as follows:

$$v_1 = \frac{V_o}{\sqrt{k e_1 k m_1}}$$

$$v_2 = \frac{V_o}{\sqrt{k e_2 k m_2}}$$

The term V_o is the velocity of the wave in homogeneous medium, and the subscripts 1 and 2 apply to the media involved. The refraction angle, θ_2 , is related to the angle of incidence, θ_1 , by the following equation:

$$\frac{\sin \theta_1}{\sin \theta_2} = \frac{v_1}{v_2}$$

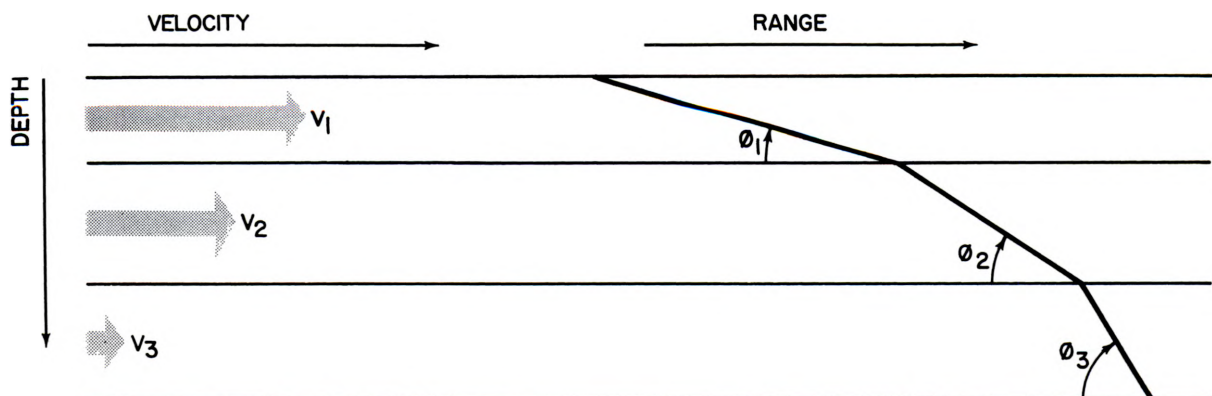


This is the basic law developed by Willebrod Snell, which states that when a wave travels obliquely from one medium into another, the ratio of the sine of the angle of incidence to the sine of the angle of refraction is the same as the ratio of the respective wave velocities in these media, and is a constant for two particular media. The angles of incidence and refraction lie on the same plane. If a plane wave is considered to be passing through three layers or strata, in each of which the sound velocity is considered to be constant, then Snell's law can be expressed as:

$$\frac{v_1}{\cos \theta_1} = \frac{v_2}{\cos \theta_2} = \frac{v_3}{\cos \theta_3} = \frac{v_n}{\cos \theta_n}$$

where v_n = velocity of sound at any point in the medium
 θ_n = angle made with the horizontal at that point.

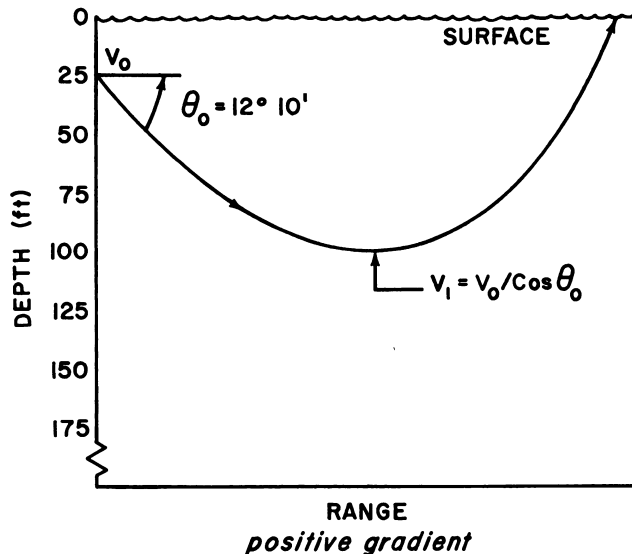
Note that the angle of inclination, θ , is the complement of the angle usually expressed in Snell's law and note also that the ray in each layer is a straight line. In practice, temperature does not change abruptly, but the gradient will normally decrease or increase in intensity, and the ray approaches a curved shape. However, at each point along the wave the velocity of the wave is still determined by Snell's equation.



NEGATIVE-POSITIVE GRADIENTS

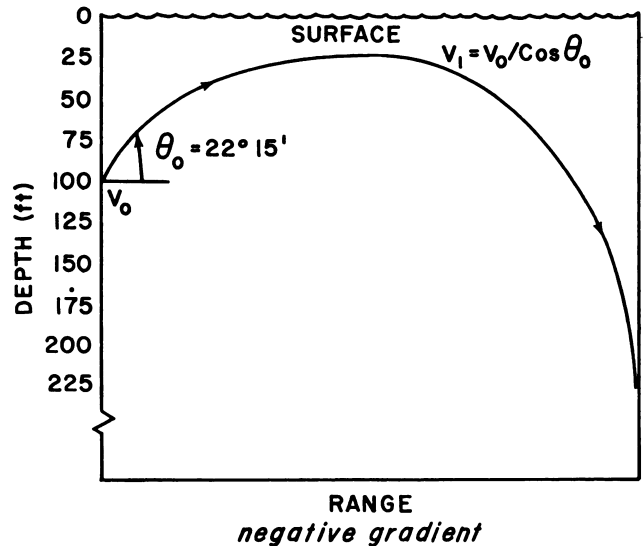
Consider a medium in which the velocity of propagation varies with change in depth, but does not vary with a change in horizontal position. Under these circumstances, a wave propagated at a depth, d , and at a velocity, v_0 , will have an initial inclination of θ_0 . When the wave enters a new medium at an inclination with the horizontal of θ_1 , then $\cos \theta_1 = v_1 \frac{\cos \theta_0}{v_0}$.

If the wave passes through a positive gradient (the velocity of sound increases with depth), $\cos \theta_1$ will increase, and θ_1 will decrease.



The wave, however, can never reach a depth where the velocity is greater than $\frac{v_0}{\cos \theta_0}$. When it approaches this optimum point, it will be refracted upward to form a concave upward curve.

Conversely, if the temperature decreases with depth (negative gradient), the velocity of the wave leaving the radiating source at an angle θ_0 above the horizontal will approach the maximum value, v_1 , at which point the wave will be diffracted downward to form a concave downward curve.



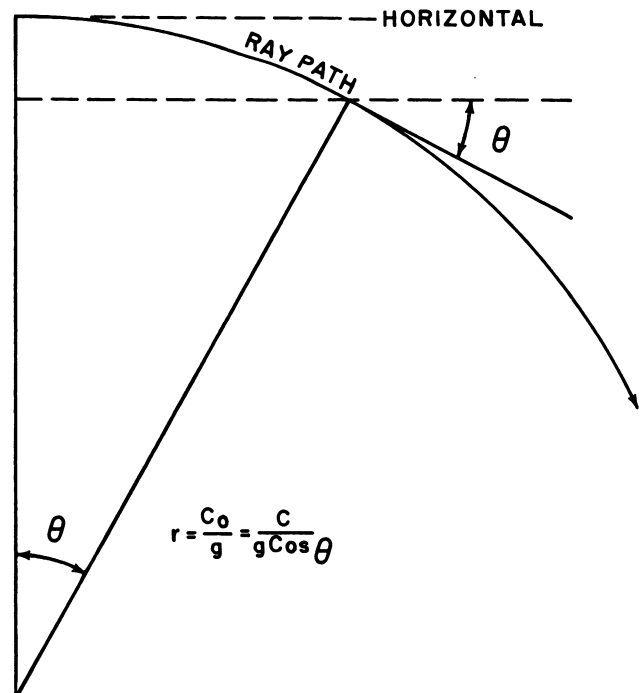
SHADOW ZONES A sound beam always bends towards the region of slower velocity; therefore, a beam going through a positive gradient (cool to warm water) will be refracted upwards. In a medium where a uniform velocity gradient exists (equal changes of velocity for equal changes in depth), the path of a ray is an arc of a circle and the radius of this circle is given by

$$r = c/g \cos \theta$$

where

g = the velocity gradient, feet per second per foot

r = radius of the arc, or radius of curvature of the path

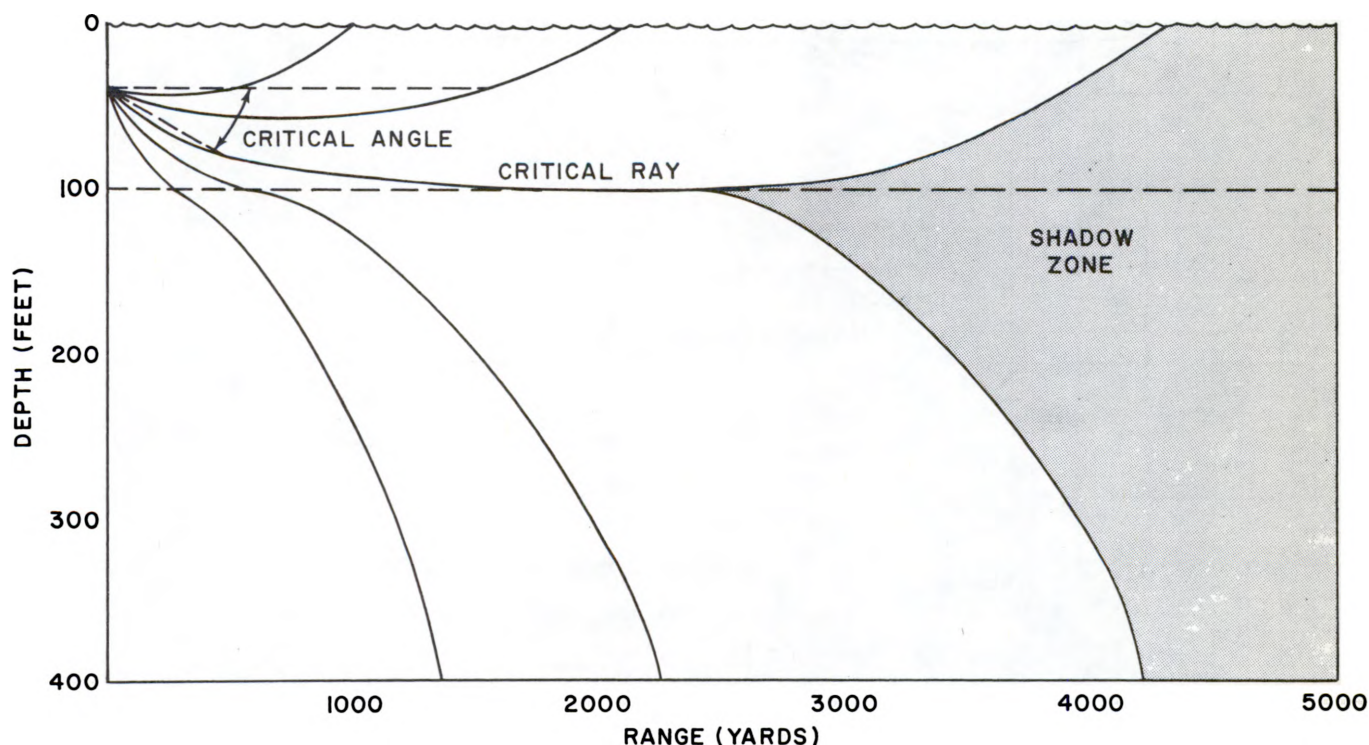


A path for a uniform negative velocity gradient is shown. Any sound wave can be subdivided into elements that travel in straight lines called rays. If a positive temperature gradient exists in an upper layer and a negative gradient exists in a layer below it, and a sound source is located in the upper layer, the sound rays will be bent upward in the upper layer and downward in the lower layer.

One ray, labeled "the critical ray" becomes horizontal at the boundary division between the isothermal layer and thermal layer. The velocity of sound is a maximum at this boundary point. One half of the critical beam bends toward the upper region at a reduced velocity and the other bends toward the lower region at a reduced velocity. The angle which the critical ray makes with the horizontal at the point of projection is called the critical angle.

All rays in the sound beam directed at an angle less than the critical angle will follow paths entirely within the isothermal layer and will be bent upward to the surface. All rays directed at an angle greater than the critical angle follow paths which strike the boundary between the isothermal layer and the thermocline at an angle. These rays pass through the boundary and are bent downward into the thermocline as shown.

No rays enter the region bounded by the two branches of the split critical ray and for this reason it is called the shadow zone. Sharp shadow zones are not fully developed because of diffraction and other effects to be discussed later, though the sound intensity in this area is quite low.



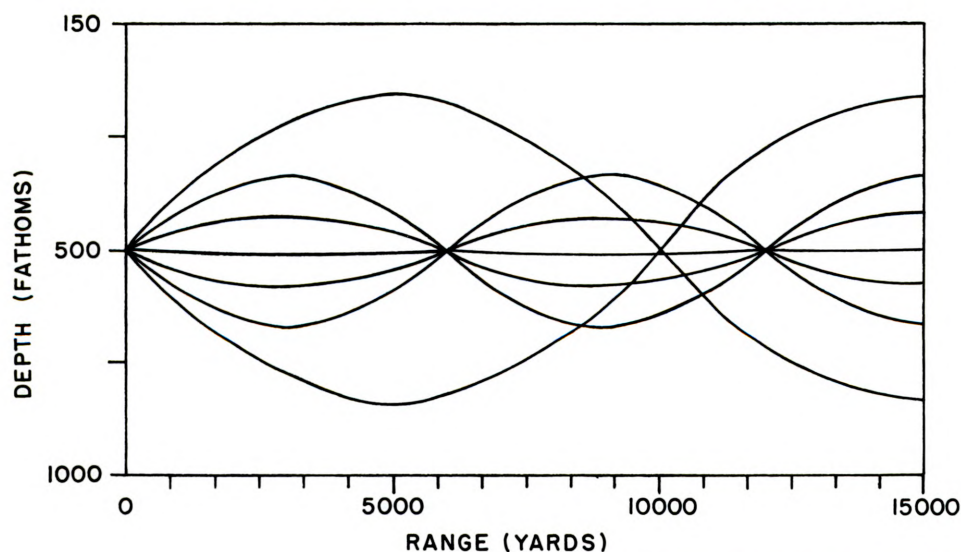
SOUND CHANNELS. The extremities of a shadow zone encompass an area in which a reduction of sound intensity occurs. This might be thought of as an area of minimum velocity of ray propagation. A sound channel is formed by a layer of water that has a negative velocity gradient overlying an adjacent layer that has a positive velocity gradient. Under these circumstances any sound signal traveling in this area is refracted back and forth so that it becomes horizontally channeled. Sound rays originating with an initial upward inclination are refracted downward, while those originating with an initial downward inclination are refracted upward. Under normal ocean conditions a sound channel with minimum velocity exists at about 500 fathoms. Rays from a sound source in this layer which make a small angle with the horizontal will follow roughly sinusoidal paths, crossing and recrossing the layer of minimum velocity. This phenomenon may be explained as follows. Consider a ray directed at an angle (θ) with the horizontal. The limiting velocity is

$$c_0 = \frac{c_{\min}}{\cos \theta}$$

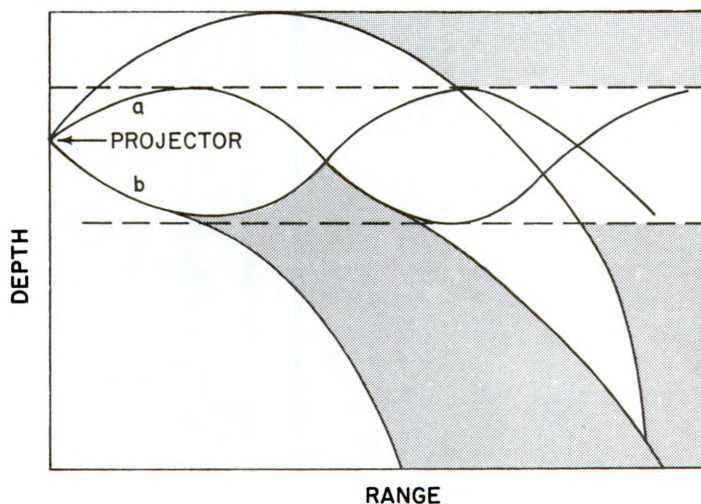
where c_{\min} is the velocity of sound in the layer of minimum velocity and c_0 is the velocity of sound in the layer above. The ray will become horizontal in this layer and

then follow a curved path back to the layer of minimum velocity. At the point where the path crosses the layer of minimum velocity, it will again make an angle θ with the horizontal, but this time the ray will be directed below the horizontal to the next layer where c_0 is the velocity of sound. The ray will be bent up to cross the layer of minimum velocity again because of the action described previously. The ray is in effect confined in a channel known as the deep sound channel between upper and lower layers where $c = c_0$. The incident angle, θ , has a limiting value above or below which the ray will strike either the surface or the bottom. However, all rays which make an initial angle θ with the horizontal less than this limiting value will be confined to the deep sound channel.

Under certain circumstances, a sound channel exists near the surface of the sea. In a surface layer with a strong positive velocity gradient the upward bending of sound rays combined with reflections from the surface will form such a channel. A sound channel in the surface layer can also result from the presence of a minimum velocity layer. Sonar ranges many times greater than normal have been observed where sound channels exist. However, the conditions that produce such sound channels in the ocean are rare and not very stable.



formation of a sound channel



velocity

The relationship between the frequency of vibration, the velocity of propagation, and the wavelength can be expressed by the formula

$$c = \lambda T$$

where c = speed or velocity of the wave

λ = wavelength

T = time duration necessary for body to complete one vibration

The period T is the reciprocal of the frequency f , and hence the wave velocity is $c = f\lambda$. For sea water at 14 degrees C, c is 150,000 centimeters per second or 4920 feet per second, and is independent of f .

The velocity of the transmission of sound in water varies with pressure, salinity, and temperature. An empirical expression for the velocity of sound in sea water at any values of temperature, salinity and depth is:

$$C = 141,000 + 421 t - 3.7 t^2 + 110 S + 0.018 d$$

where C = velocity of transmission in centimeters per second

t = temperature of the water in degrees centigrade

S = salinity of water in parts per thousand

d = depth below surface in centimeters

The wave velocity is determined completely by the properties of the transmitting medium and does not depend upon either the frequency of the source or the wavelength.

For example, temperature ordinarily affects density to a greater degree than it affects the elasticity modules. Thus, the higher the temperature of the medium the lower the density, and the higher the velocity. Of the three factors (temperature, pressure, and salinity) that

Velocity is the time rate of change of wave position. Because velocity is a vector quantity, the determination of velocity requires that a scalar magnitude expressed in wavelengths divided by time be determined together with a direction relative to a frame of reference. When sound waves are propagated within a liquid, the velocity of the wave is determined by the elasticity and the density of the medium, by the temperature of the medium, and, to a limited degree, by the salinity and pressure characteristics of the medium.

The expression of this velocity is determined by comparing the pressure and velocity of the particles within a condensation period with the same characteristics in an undisturbed medium, and evaluating the force which produces the change of motion.

The velocity of sound in a liquid is given by the equation

$$c = \frac{\sqrt{E}}{d}$$

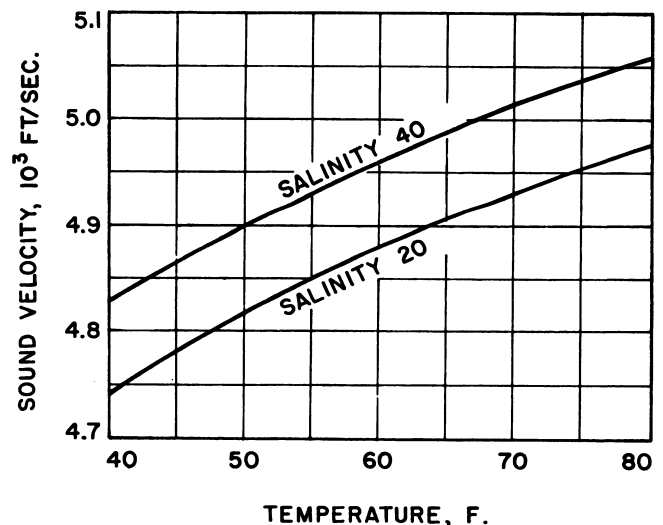
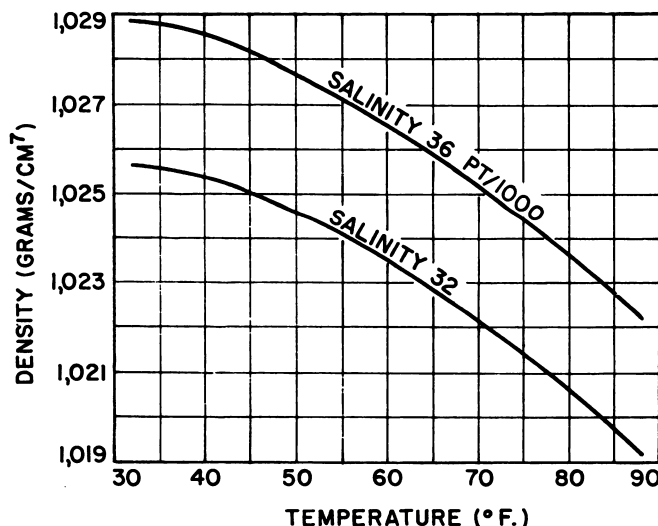
where c = velocity in centimeters per second

E = bulk modulus of elasticity of the liquid in dynes per square centimeter

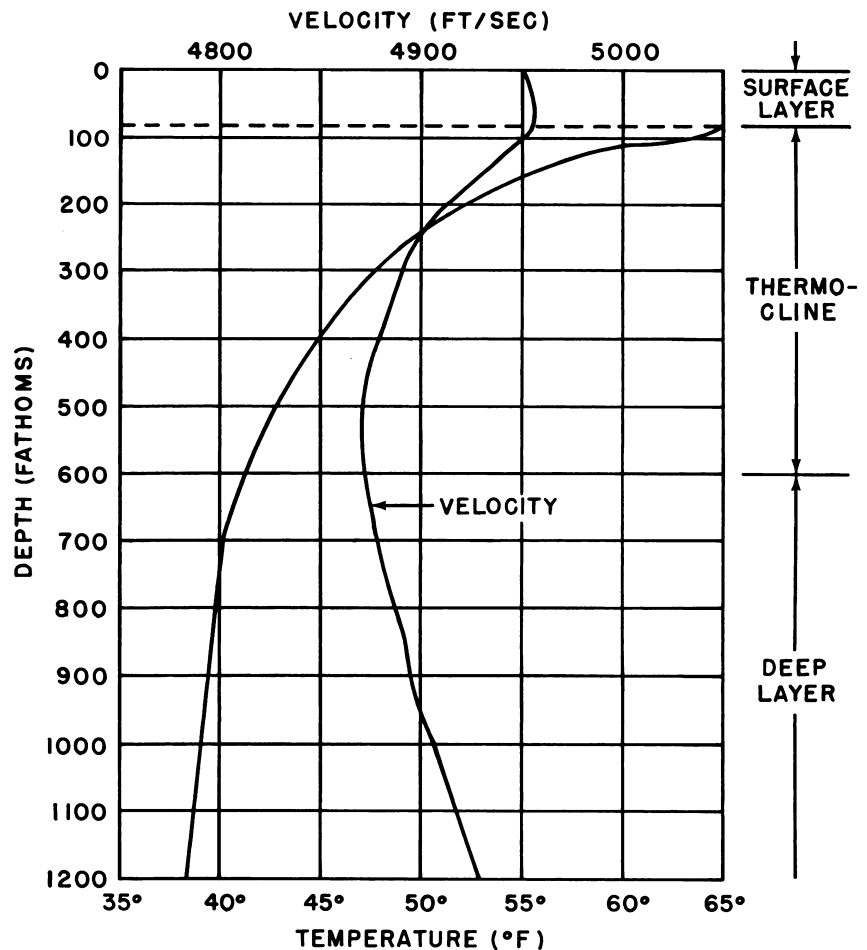
d = density in grams per cubic centimeter.

Because the density of sea water varies with the pressure at transmission levels, the salt content of the medium, and the temperature of the liquid, the effect of these three parameters on density must be considered. Regardless of the factors that determine density, velocity increases with an increase in the density coefficient of the medium.

affect the elasticity modules and density of the medium, temperature is by far the most important in sound transmission. Note in the illustration that at constant salinity the density of the medium changes at a variable rate, and thus the velocity increases with change in temperature at a variable rate.



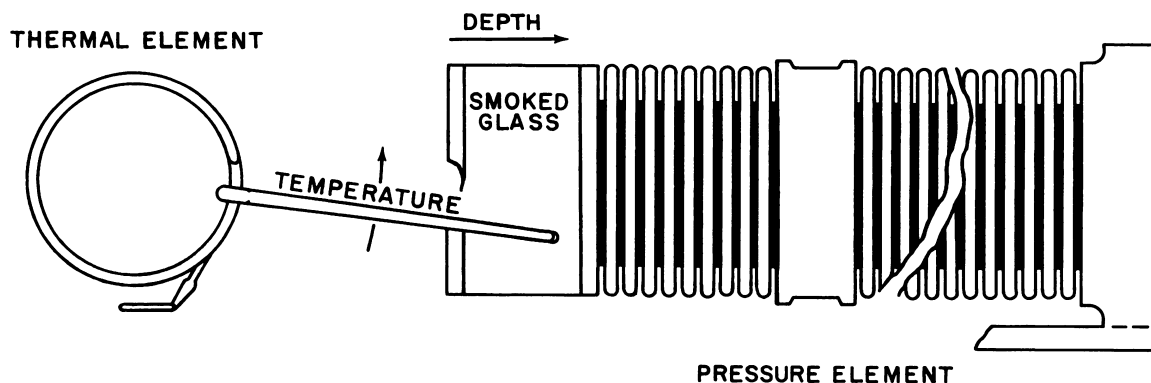
TEMPERATURE GRADIENTS. The oceans may be considered as consisting of strata, in any one of which the same temperature exists over a large horizontal distance. The first stratum is commonly called the surface layer, and it is in this layer that wide distributions of temperature versus depth occur. Surface conditions, wind action, external temperatures, tidal motion are but a few of the factors that cause a wide divergence in temperature readings. Immediately below the surface layer, thermal conditions are normally stable. This region is called the thermocline stratum, and the temperature in this region decreases progressively with depth to a value of 40° F at a depth of 600 fathoms. Below this depth, there exists a deep stratum in which the temperature decreases very slowly with increasing depth. Strata in which the temperature is uniform are called isothermal layers. Negative gradients exist in strata in which the temperature decreases with depth. Positive gradients describe conditions in which the temperature increases with depth. A layer in which the temperature decreases very rapidly, particularly if it is immediately beneath an isothermal layer, is commonly called a thermocline.

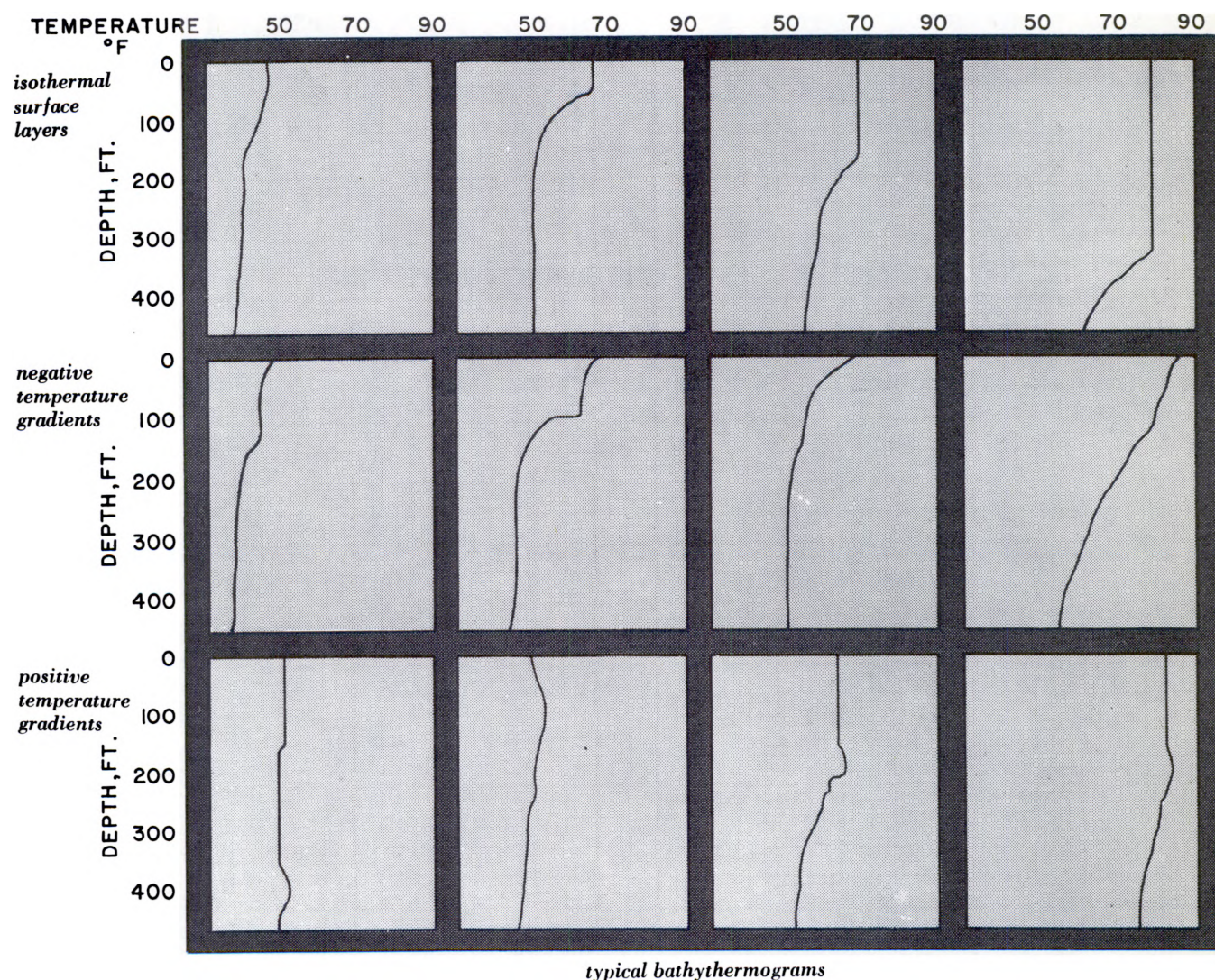
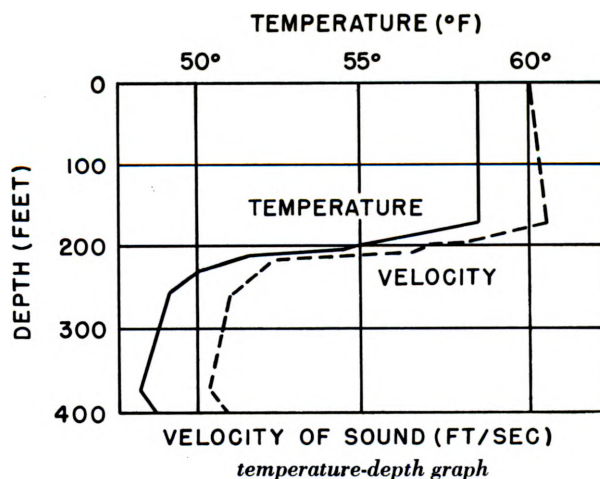
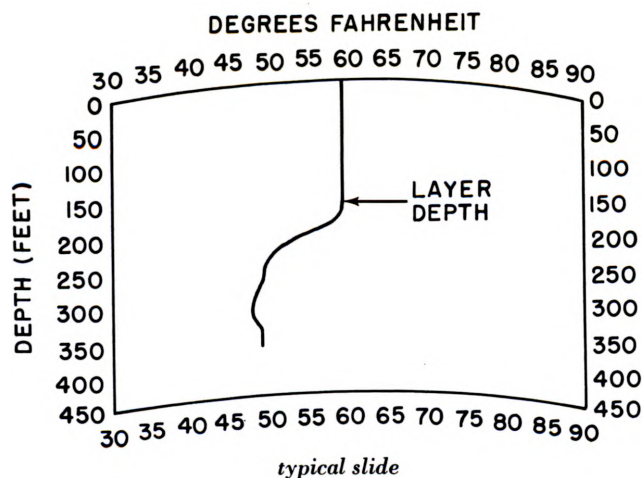


BATHYTHERMOGRAPHS. Because of the importance of thermal layers in velocity of sound propagation, they have been the subject of considerable study. The instrument used for this purpose is called a bathythermograph (frequently referred to as a BT). As it is lowered into the sea, a graph of temperature as a function of depth is automatically plotted.

The instrument consists of a stylus that is actuated horizontally by a pressure-sensitive element, and perpendicularly by a temperature-sensitive element. As

the instrument is lowered, the stylus is activated by the expansion or contraction of a thermal element in a copper thermometer tube. The increasing hydrostatic pressure compresses a bellows system, which moves a smoked glass slide at right angles to the stylus. The stylus is driven by the thermal element. Values of temperature as a function of depth can then be determined from the slide by reading points of the plot on an optically superimposed scale.





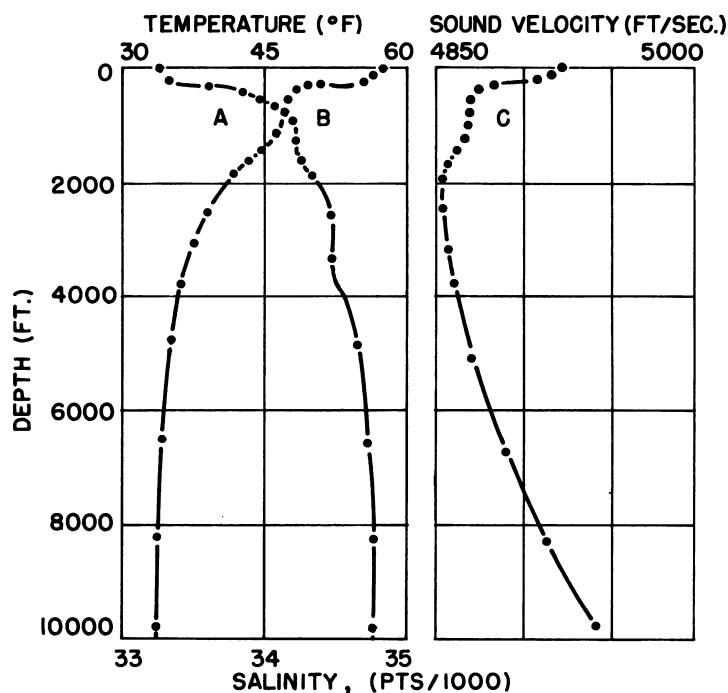
PRESSURE

The velocity of sound in the atmosphere is unaffected by changes in barometric pressure because the density and pressure changes take place in the same proportions. However, when modulus of elasticity computations are made in sea water it becomes necessary to determine how the volume of the liquid is affected by changes in pressure (p). Increase of pressure with depth causes an increase in the speed of transmission of 1.82 feet

per second per 100 feet of depth. The pressure effect is important only if both the temperature and salinity are constant. At greater depths, temperature and salinity are fairly constant, and the pressure effect dominates. Average temperature changes to a depth of 2500 feet can usually neutralize the effect of salinity and pressure. At greater depths the pressure effect is primary and sound velocity increases.

SALINITY

As the salinity of the medium increases, both its density and the velocity of sound through it increases. An increase in salinity of 1 part in 1000 increases the velocity of sound 4.27 ft./sec. In most cases the effect of salinity can be neglected, because salinity is comparatively constant except in areas where fresh water sources empty into the sea.



transmission characteristics

Besides these phenomena which affect the velocity of sound, consideration must be given to those affecting the intensity of sound. The intensity of sound decreases as the distance from the source increases. This decrease in intensity is known as propagation loss and is

subdivided into two different types. One loss is due to the geometry of the spherical wave and is known as spreading loss. The other is due to the physical properties of the wave and the medium through which it is propagating and is known as attenuation.

spreading loss

The ray path used for analysis is the radius of a spherical wave front. Since the area of this wave front increases as it moves further away from the source, the intensity of the sound distributed over it decreases. The reduction in intensity due to increased wave front area is known as spreading loss. Consider a small, spherical sound source in a limitless, homogenous, lossless medium. Such a source radiates sound equally in all directions. If the power radiated by the sphere is $p(t)$ and we take a sphere of radius R with its center at the center of the radiator, the area of the sphere will be $4\pi R^2$ and the power passing through any unit area on the surface of the sphere is the intensity of the sound, I . Since R is equal to the velocity of propagation (c) times the time (t), the equation for I is therefore:

$$I = \frac{p(t - \frac{R}{c})}{4\pi R^2} = \frac{p(t - \frac{cT}{c})}{4\pi (cT)^2} = \frac{p(t-T)}{4\pi (cT)^2}$$

If the power taken is the average power (P) then

$$I = \frac{P}{4\pi R^2} = \frac{P}{4\pi (cT)^2}$$

The intensities of two points at distances from the sound source of R_1 and R_2 are related by:

$$\frac{I_1}{I_2} = \frac{R_2^2}{R_1^2} = \frac{T_2^2}{T_1^2}$$

If $I_0 = I_2$, and R_2 is chosen as a unit distance, then the intensity (I) at any distance (R) may be written as

$$I = I_0/R^2$$

Intensity is inversely proportional to the distance squared, which is the inverse-square law of spreading.

Actual sound radiators are seldom if ever small spheres and the sea is far from the limitless, homogenous, lossless medium postulated above. However, along a given ray path, sound intensity does diminish in accordance with the inverse-square law with the exception of wave transmission in the deep sound channel.

The spreading loss in the deep sound channel does not follow the inverse square law because the sound is effectively confined to a horizontal layer. The reduction in intensity is directly proportional to distance. This enormous reduction in the spreading loss makes it possible to detect sounds originating in the deep sound layer at great distances. In a Sound Fixing And Ranging (SOFAR) system, a ship in distress releases a small hydrostatically actuated bomb set to explode in the deep sound channel. The resulting explosion is heard by strategically located permanent listening stations. The vessel which dropped the bomb is located from the differences in time of arrival of the sound at the different listening stations. The useful range from the signal source to the monitor stations can exceed 3000 miles.

attenuation

The reduction in intensity caused by the combined effects of scattering and absorption is called attenuation. Scattering is the diffusion of sound waves in a random manner. It is caused by the presence of foreign bodies (marine organisms, etc.) and by inhomogeneities in the medium (random variations in temperature, salinity and density).

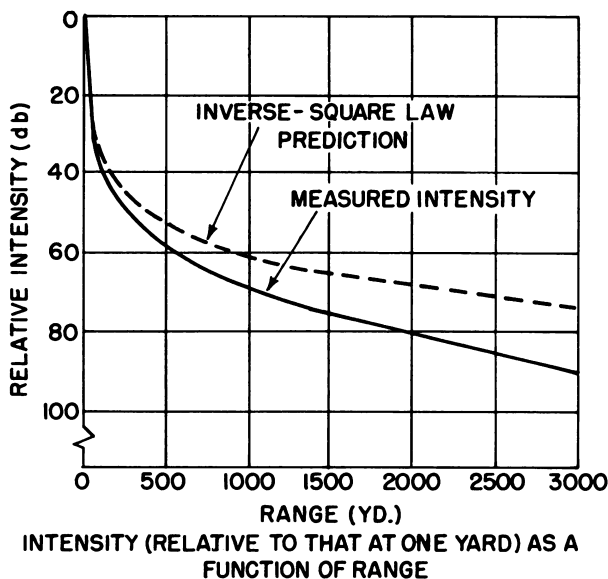
Absorption is the conversion of a portion of the acoustic energy of a sound pulse into heat energy. It is caused by the viscous friction between the water molecules as the sound pulse propagates. Mathematically, the effect of absorption on intensity is expressed:

$$I = \frac{I_0}{R^2} e^{-aR}$$

where R = range in kiloyards

a = coefficient of absorption and has a value of 0.0023 f^2 , if f is in kilocycles per second.

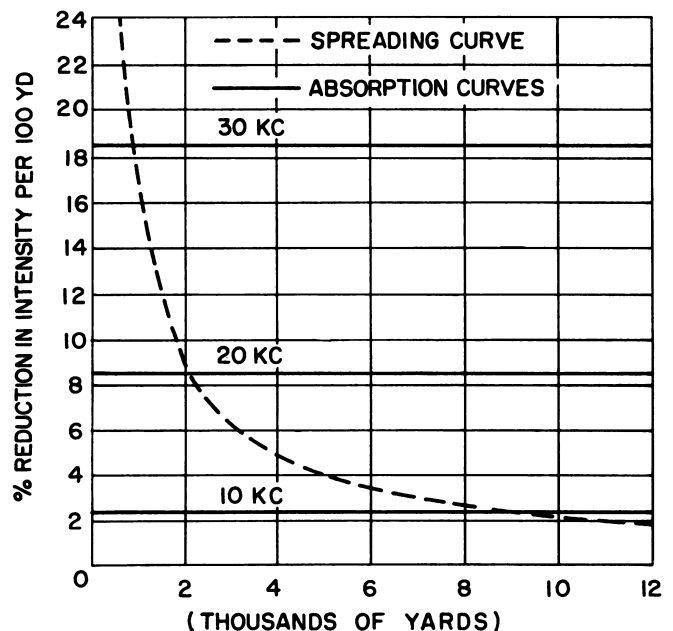
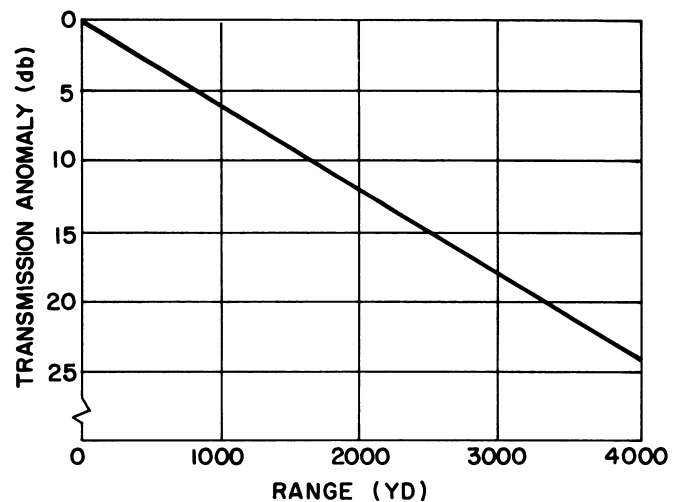
($f < 50$ kc)



effect of frequency on range

The loss of intensity of a sound wave due to spreading is a geometrical phenomenon and is independent of frequency. As range increases, the percentage of intensity lost for a given distance traveled becomes increasingly less. However, the loss of intensity caused by absorption is independent of range and increases as the square of the frequency. Useful range is severely restricted by absorption at higher frequencies. For this reason, where long-range operation of sonar equipment is desired, the lower the frequency used the better. Older echo-ranging equipments usually operated at either 26 or 30 kilocycles and seldom obtained ranges beyond 2000 yards. Such ranges are inadequate against high-speed submarines; consequently, the trend in modern equipment is toward lower frequencies for increased surveillance probabilities.

Attenuation increases with range. As a result, the amount of acoustic energy deviates more and more from the value predicted by the inverse-square law as the distance from the source is increased. The discrepancy between the measured and the calculated curve is not appreciable for ranges of less than 200 yards. The discrepancy between the measured intensity and that predicted by the inverse-square law is called the transmission anomaly and is expressed in decibels. The transmission anomaly in decibels plotted as a function of range for the conditions just illustrated is a straight line. The transmission anomaly is a linear function of range, unless the transmission conditions change with range.



reflection

Any body immersed in water acts as a reflector of underwater sound. Reflections from the hulls of submerged submarines are of particular interest since they cannot readily be detected by other means. An immersed object that is large in comparison with the wavelength of sound striking it will cast an acoustic shadow, i.e., the intercepted sound power will be reflected as secondary sound transmission, called an echo. The amount of sound power reflected by the body is determined by the area of the shadow cast by the body upon a plane perpendicular to the direction of the sound rays. This area, referred to as the effective target area (A_T), is equal to $\pi d^2/4$ for a spherical body. For a body that is not spherical, the area of its acoustic shadow depends on the position of the body in the sound field. If W is the total power intercepted by the target, and F is the amount of energy flow in watts per unit area at the target,

$$W = F A_T.$$

All intercepted energy is reflected as soundwaves if the target is a perfect reflector, otherwise part of the inter-

cepted energy is converted to heat. A large flat body acts as a mirror; all intercepted sound is reflected in a single direction. The surface of the sea is perhaps the closest approach to a mirror-like reflector encountered in underwater sound propagation, but the motion of the surface is usually sufficient to give it properties different from those of a mirror. The ocean contains many small particles, such as air bubbles, fish, plants and other marine life, whose dimensions are small in comparison with the wavelength of the sound. The effective target area of such particles is less than their actual area by a ratio approximately equal to $\pi d/4\lambda$, where d is the diameter of the particle, and λ is the wavelength of the sound.

Gas bubbles are normally not as widely distributed in the ocean as solid particles are. Gas bubbles may occur near the surface when waves are breaking on the surface or they may occur in the wakes of ships. Therefore, bubbles usually occur in large groups. Such localized groups of bubbles may have considerable effect on sound transmission through the ocean.

BOTTOM REFLECTION

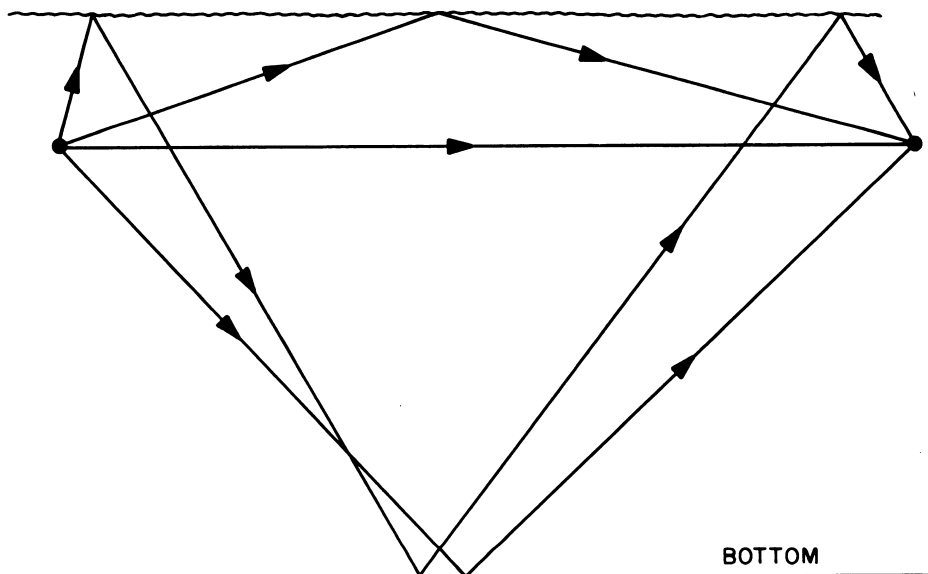
The effects of sound reflected and scattered by the ocean bed is usually negligible in deep water since the equipment used to emit the sound waves is generally near the surface. However, in shallow waters bottom reflection can be the limiting factor on the range from which intelligible echoes can be obtained. When the reflection is from the surface, the intensity and direction of the scattered sound is determined by the state of the sea. The reflections from the bottom, however, are dependent on the nature of the ocean bed. Ocean beds fall into three categories:

- 1) rock
- 2) sand
- 3) mud

Rock and sand bottoms are good reflectors and can affect sound transmission in shallow waters. Mud, on the other hand, is such a poor reflector of sound that its effects are negligible.

SURFACE REFLECTION

The change in acoustic impedance in passing from water into air is so large that the surface is almost a perfect reflector for underwater sound. If the surface were perfectly calm, it would behave like a flat mirror. This condition, however, is practically nonexistent in the ocean and the surface presents a continuously changing, uneven boundary, that randomly reflects sound waves in all directions regardless of their point of origin. The motion of the ocean surface under normal conditions will result in random scattering of sound at frequencies above about 10 kc (wavelengths of less than about 7 inches). In addition, the breaking of waves causes a considerable population of small bubbles in the region near the surface. Consequently, the surface and the water bordering it is a region of random scattering. Scattering of this sort from this region is called surface reverberation.



reverberation

We have already seen that both the surface and the bottom of the ocean reflect acoustic energy. When a portion of this scattered energy is reflected back to its point of origin, it is known either as surface reverberation or as bottom reverberation. Much of the energy reaching the starting point is a result of reflection in its travel.

In addition to those at its boundary surfaces, impedance discontinuities are to be found throughout the volume of the ocean itself. These may be due to a variety of causes such as air bubbles generated by wave action and gas bubbles from decaying vegetation or living organisms. Frequently these reflectors are of considerable size, as in the case of masses of kelp or of schools of fish. Such large reflectors are to be found at intervals and are usually reported as targets. There are so many minute reflectors in the ocean that no sample

is free of them. It is these small reflectors, or scatterers, which are responsible for what is known as volume reverberation. Scatterers are understood to have dimensions which are small when compared with those of a signal wave. Each scatterer may be considered to act as a point source of secondary radiation.

Although scatterers are found in all parts of the sea, they are not uniformly distributed. They are sometimes found concentrated in layers of varying area and depth, occasionally moving in a daily cycle. Under some conditions, the reverberation resulting from a layer having a high concentration of scatterers exhibits the same general characteristics as bottom reverberation.

Reverberation resulting from a pulse of acoustic energy varies with the range at which the pulse is reflected. Reverberation suffers from propagation losses in water as do all forms of sound transmission.

Although reverberation cannot be eliminated, it can be minimized if its nature is understood.

effect of thermal microstructure

The amplitude of the signal received from a fixed source of constant amplitude fluctuates in a random manner. Similarly, the intensity of successive reflections from a fixed target varies widely when using echo-ranging gear. The cause of these fluctuations was long a mystery. It is now believed that they are caused by the thermal microstructure of the sea, i.e., temperature differences within a layer due to incomplete mixing in the layer. The temperature changes associated with this thermal microstructure occur over distances in any direction as short as a yard and are randomly distributed. As a result, sound rays over a wide angle from a given source follow irregular paths to a receiving de-

vice at a given point. The number of rays reaching the receiving device and hence the amplitude of the received signal will vary from moment to moment as the thermal microstructure between the two points changes with the movement of the water and changes in temperature. One of the consequences of this phenomenon is that a single observation will seldom yield results that agree with those predicted from the application of the principles set forth in the preceding sections. However, the average of a number of observations, where random fluctuations tend to cancel, will agree with predicted results. If it were possible to determine the thermal microstructure accurately, it would be possible, but not practical, to predict the behavior of sound with considerably greater precision.

doppler effect

Because of the varying effects of refraction and attenuation in the ocean, the relative motion between target and seeker is difficult to establish accurately. For this reason, the Doppler shift in sonar systems is determined by comparing two echoes. The first echo, or reference frequency, is that returned from the sea, which is practically motionless. This is the reverberation frequency discussed previously. The second echo is that reflected from the target.

If the echo frequency is greater than the reverberation frequency, the target is moving toward the sound transmitter. This is known as up Doppler. If the echo frequency is less than the reverberation frequency the target is moving away from the sound transmitter. This is known as down Doppler.

The relationship is:

$$f_e = f \left(1 \pm \frac{2v}{c} \right)$$

where f_e = echo frequency

v = relative velocity between target and seeker

c = the local speed of sound.

SOUND SOURCES AND NOISE

Sound present in the waters of the ocean results from natural phenomena, from the activities of marine life, and from the activities of man. Sounds are often made up of many components, the magnitudes of which vary

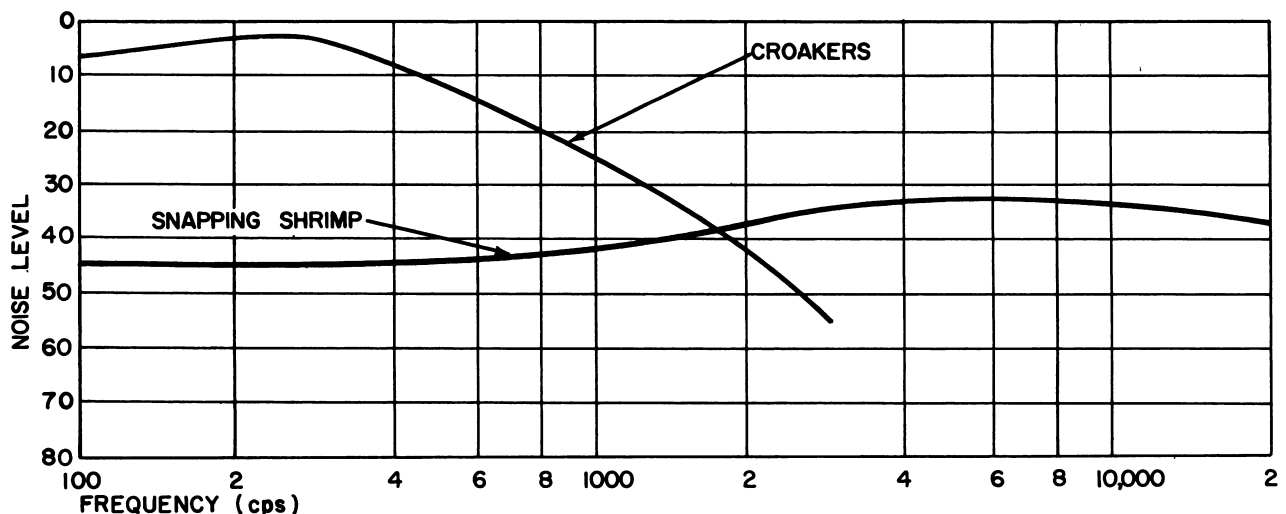
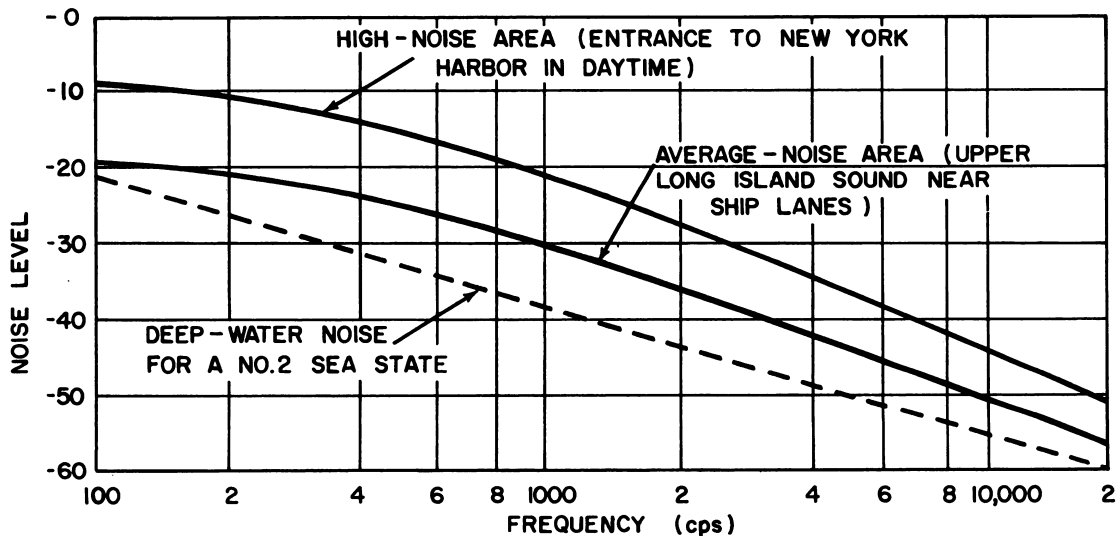
with time in a random manner independently of each other. If intelligible characteristics are not purposely imparted to these sounds for the purpose of establishing communication, they are referred to as noise.

natural phenomena

There are many sources of natural noise in the ocean. Waves at the surface are the largest source. The intensity of surface wave noise varies with the state of the sea. Surface noise has a frequency range of approximately 0.1 kc to 50 kc. Lesser natural sources of noise include: molecular agitation in the ocean

(causes thermal noise) and external natural phenomena such as storms, earthquakes, etc.

Noise of this type presents a problem in the operational use of sonar because it interferes with normal echoes, obscuring targets or being mistaken for targets. It can also attract or confuse acoustic torpedoes or sonar-type missiles.



MARINE LIFE NOISE

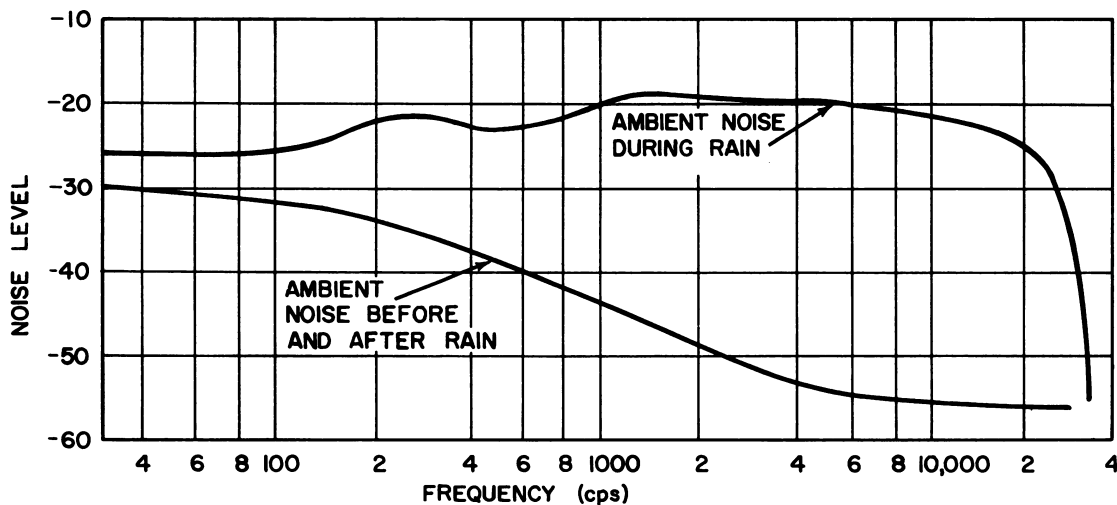
Although several different organisms contribute to the noise in the ocean (shrimp, croakers, etc.), by far the most important contributors are the snapping shrimp. Snapping shrimp produce noise by snapping their claws. They are found in large beds at the bottom of the ocean. A single snap by a shrimp produces a pulse that lasts less than a millisecond, but the entire shrimp bed produces a continuous noise.

Small fish called croakers also contribute to marine noise. A croaker or drumfish is equipped with a gas-filled bladder on which it beats with a vibrating muscle at a rate of about seven blows per second for two seconds. This is repeated at intervals of five to ten seconds. Whenever croakers come close to a hydrophone, the sounds, reproduced by a loudspeaker, are

similar to those of a woodpecker. The sonic output of a single croaker is not large. However, croakers often travel in large numbers and the sound coming from a loudspeaker becomes a roar in which drum rolls are intermittently detected.

This phenomenon is most noticeable for an hour or so after sunset, at which time it can seriously impair reception of frequencies below 2 kc. By using a filter in the sonar receiver to block frequencies below 2 kc, ship or torpedo noises at high frequencies are easier to detect.

Other marine animals produce noises, but so far these noises do not significantly affect acoustic homing torpedoes or sonar. These sources include porpoises (which talk or bark), whales, fish with noise-producing bladders, mackerel (which click their teeth), trigger fish (which click joints on their fin spines), relatives of the snapping shrimp, etc.



man made noise

Much of the noise present in the sea is caused by the activities of man. The most common and pertinent of these sounds are those made by ships. The following discussion therefore is concerned with the noise characteristics of various man-made sea vehicles. Any vehicle moving through the ocean produces noise, the frequency and intensity of which is dependent on its size, shape, speed, etc. The noise from water vehicles is classified as hydrodynamic noise, hydroelastic noise, machinery noise, and artificial noise.

HYDRODYNAMIC NOISE

The sources of hydrodynamic noise are natural physical forces and cavitation effects.

Cavitation is produced on a body that has motion relative to the water because the pressure at various points is reduced below the vapor pressure of the water. These pressure reductions are caused by the shape of the body. In general, a body without appendages has less cavitation than a body with appendages. If such additions are

necessary, they must be designed to eliminate as much cavitation as possible. Besides forming on the hull and various structural members of a sea vehicle, cavitation can also take place at the propeller.

The noise level produced by cavitation falls off at the rate of 6 db per octave at frequencies up to 100 kc. However, above this frequency the level remains fairly constant, with the result that almost all other noises in the ultrasonic region are masked.

Since all torpedo homing systems and many sonar systems operate in the ultrasonic region, cavitation noise is a serious problem. Torpedoes generally home on the cavitation noise produced by ships, and any cavitation noise produced by the torpedo interferes with target noise.

Because the speed at which a vehicle with a given cavitation number can operate without cavitating increases as the ambient pressure is increased, some acoustic torpedoes are designed to search and attack from depths known to be below the cavitating depth. In the same way, a submarine commander, when under attack, will attempt to dive to depths at which he can attain a relatively high speed without producing cavitation.

HYDROELASTIC NOISE

Propellers and appendages on bodies being propelled through the water produce a singing noise at certain combinations of speed and depth. The hum of telephone wires under a steady wind force is an example of singing caused by the unsteady flow of air in the wake of the wire. The unsteady flow is caused by vortices that are alternately shed from the top and bottom of the part. Although the frequency is in the low audio region, it usually generates harmonics which affect the ultrasonic region. Besides confusing attacking torpedoes, hydroelastic noise can result in detection of the torpedo by the vessel under attack.

MACHINERY NOISE

The dominant source of machinery noise in an underwater vehicle is its power plant and power distribution system which supplies power to the other machinery on the vehicle such as compressors, generators, propellers etc. Machinery noise is always present unless the vehicle is propelled by impulse or by the movement of the water and has no moving parts. Machinery noise is kept to a minimum by acoustically isolating the various moving mechanical components.

The gearing connecting the propellers is an important source of machinery noise. If the engine runs at a relatively low speed, as is the case with a reciprocating heat engine, gears may not be required between the engine and the propeller. High-speed power sources, however, such as some types of electrical motors, usually require reduction gears to the propeller.

The frequency of the explosions in the cylinders of a reciprocating engine is not likely to be a source of ultrasonic noise but might be an important source of low-frequency sonic noise. A more crucial source of noise in the reciprocating engine is the clatter of the valves opening and closing.

In gas turbines the noise generated is in the ultrasonic region at a radian frequency equal to the angular velocity of the turbine buckets.

Noise in the ultrasonic region is extremely important in sonar and acoustic torpedo performance. The noise produced by auxiliary units, such as pumps, generators, servos, and even relays, is often more significant than the power-plant noise. The large masses involved in the power plant usually keep noise frequencies relatively low. For this reason a small relay may interfere more with the operation of a torpedo than an electric motor that generates several horsepower. Small, high-speed servomotors, however, may be serious sources of ultrasonic noise.

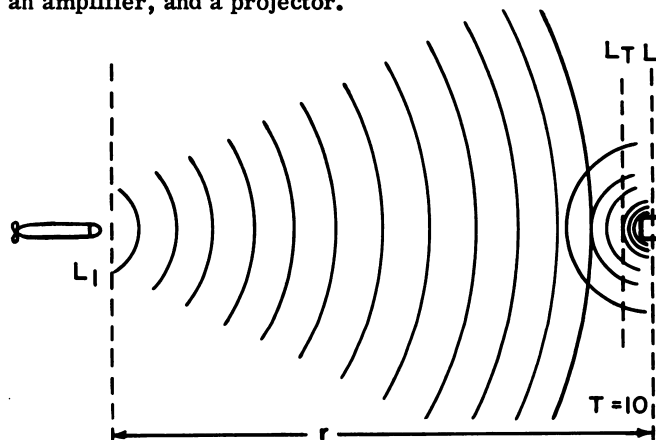
ARTIFICIAL TARGETS

Artificial targets simulate the noise or echo of ships of known characteristics. They are used for the experimental testing of acoustic torpedoes and sonar systems. Such targets are used also as countermeasures for decoying acoustic torpedoes or confusing enemy sonar equipment. The transducers used as sound projectors in these targets are excited by electronic gear and can be adjusted to simulate the noise or echo levels of any desired ship.

THE SIMPLE NOISE TARGET consists basically of a floating buoy from which an omnidirectional transducer is suspended by cable to the target depth desired. The transducer is actuated electronically and emits sound intensities at predetermined frequencies at measured intervals.

An artificial noise target must simulate both the noise frequencies of the target and the intensity with which these frequencies are present in the target. Since intensity will vary with range, to obtain a correct elevation of sonar equipment it is essential to know the exact location of the artificial target with respect to the equipment under test. For this reason, target buoys are often equipped with radar reflectors which enable the exact range of the artificial targets to be found by radar at the time of testing.

ECHO-REPEATER TARGET. Echo-repeater targets simulate the reflecting characteristics of the hulls of ships. They are used to test active sonar systems and active acoustic torpedoes. The artificial target radiates a signal which simulates the reflection of a specified target. The signal level radiated is determined by the target range and aspect being simulated. The echo repeater shown here consists of a receiving hydrophone, an amplifier, and a projector.

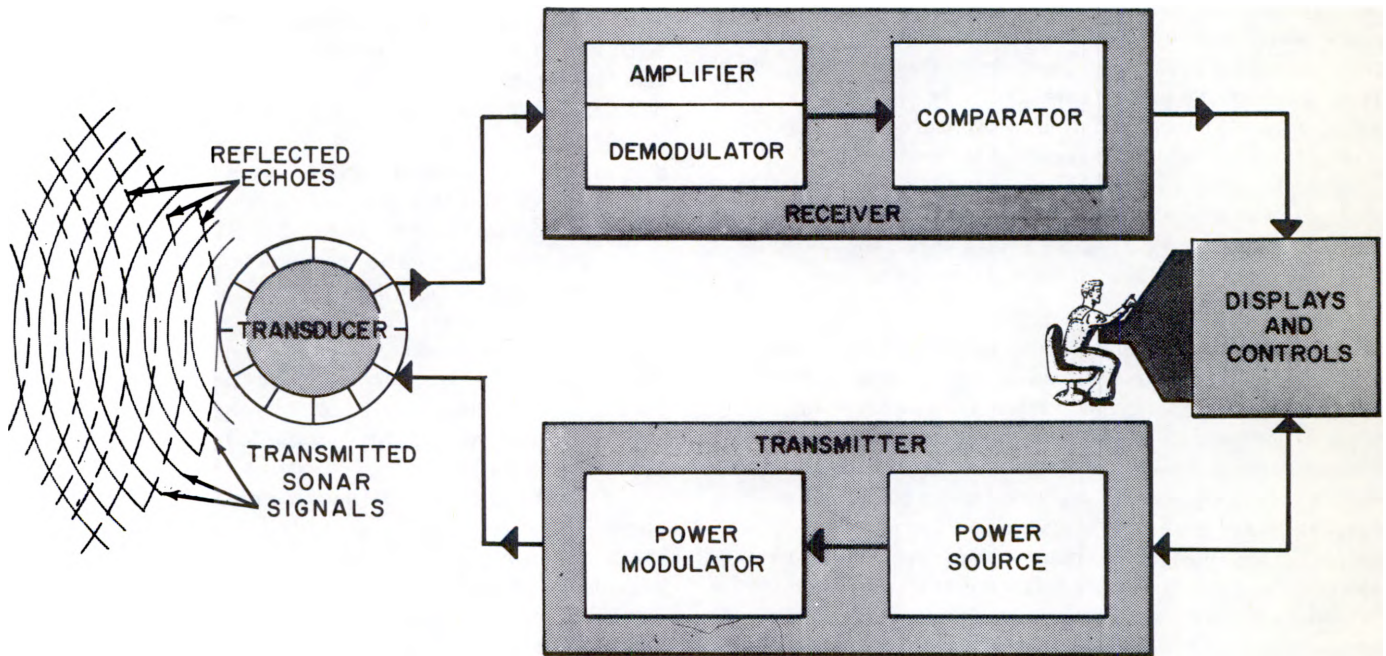


ECHO-REPEATER DOPPLERIZERS. The simplest method of simulating the Doppler effect that is produced by a moving target is to tow an echo repeater at a speed corresponding to that of the target being simulated. Because of the inconvenience of towing targets, electronic means were developed to simulate the Doppler shift by causing the signal transmitted from the echo repeater to differ in frequency from the signal it received by the amount which would result from the aspect and speed of the target being simulated. An echo-repeater Dopplerizer consists of a receiving hydrophone, a dual superheterodyne circuit that converts the input frequency to the desired output frequency, a transmitter, and a projector.

COUNTERMEASURES. Echo-repeater buoys can be dropped by a vessel to lure enemy torpedoes or to confuse enemy sonar with false targets. Echo-repeater Dopplerizers can also be used on board a ship or submarine to transmit false signals to mask their own echoes and to confuse enemy sonar.

BASIC SONAR SYSTEMS

The various types of sonar used in weapons systems are divided into three basic categories:
ECHO-RANGING, LISTENING, and COMMUNICATIONS SYSTEMS.



ECHO-RANGING SYSTEMS are used to detect the presence of ships, analyze shoreline and bottom characteristics, and determine bottom depths. The basic echo-ranging system consists of a transmitter, one or more transducers, a receiver, and displays and controls of various sorts. The transmitter has a source of power and a means for modulating the basic power in a manner suitable for the particular application. In general, the basic transmitted power is an ultrasonic frequency modulated as single pulses at a pulse repetition frequency (prf) dependent on the system range. The transmitter power is converted from electrical to acoustic energy by means of the transducer for transmission of the sonar signal into the water. Received signals or echoes are reconverted by the same transducer (or by a different transducer in some sonar systems) from acoustic back to electrical energy. The received signal from the transducer is processed in a receiver, where various amplification, demodulation, and comparison operations are performed. The output of the receiver, in the form for proper presentation of target data, is fed to display and control units. The displays include aural types, such as loudspeakers and earphones, and visual types, including oscilloscopes, recorders, and indicator lamps.

LISTENING SYSTEMS utilize only the receiving elements of the echo-ranging system. These are the transducer, receiver, displays and controls. The transducer in a listening system is generally called a hydrophone. An echo-ranging system would be used on a vessel for passive listening only under circumstances in which it was not tactically desirable to make the presence of the vessel known by transmitting. In this case, the transmitter would be turned off.

COMMUNICATIONS SYSTEMS are basically similar to the echo-ranging systems, although generally simpler in design. Transmission consists of pulsed code or voice-modulated ultrasonic power. The reception consists of demodulation and amplification stages. A comparator is not needed, except to keep the transducer trained to maximize reception. The display portion consists only of the aural types. Within each category are many varieties. The several types of echo-ranging systems are divided basically into searchlight and scanning sonar systems. Many echo-ranging systems also are capable of both direct listening and communications.

transducers

A device for converting one form of energy to another is a transducer. In sonar the acoustic energy of the sound waves is converted to mechanical energy in the form of oscillation of the molecules of the medium through which the sound travels. These oscillations cause a synchronous variation in pressure of the medium. The signals generated or received by the electronic circuits in sonar equipment are in the form of electrical energy. The sonar transducer acts as the link between the water and the electronic circuits of the sonar equipment and converts the electrical energy to acoustical energy and vice versa.

There are physical phenomena which exhibit the ability to change electrical to mechanical energy and mechanical to electrical energy, and which are employed in sonar transducers. These are the electrostrictive, the piezoelectric and the magnetostrictive effects. Materials exhibiting electrostrictive and piezoelectric properties are generally of crystalline or ceramic nature. They change dimensions when subjected to an electric field and develop a voltage potential between two opposite faces when mechanically stressed. Materials exhibiting the magnetostrictive effect change dimensions when subjected to a magnetic field and change magnetic permeability when subjected to mechanical forces. This change in permeability changes the intensity of any magnetic field in the presence of the material.

Variation in the strength of the electric or magnetic field at a frequency in the acoustic spectrum will cause acoustic waves to be generated by such material. When the transducers are placed in water, such acoustic waves can then be propagated from the material to the water. When sound waves impinge upon such material after propagation from the water, the mechanical stresses resulting from the pressure variations will cause variations in the field strength (electric or magnetic) which in turn generate electric signals in suitable circuits.

CRYSTAL TRANSDUCERS

The piezoelectric effect, which is a form of electrostriction, is the electric polarization produced by mechanical strain in certain classes of crystals. The polarization and the electric potential induced by it is proportional to the strain and changes sign with it. In the converse piezoelectric effect, an electric potential across the crystal face produces a mechanical deformation proportional to the induced electric polarization and of the same sign. The magnitude of this effect varies for crystals of different materials, and for the different axes of the crystal.

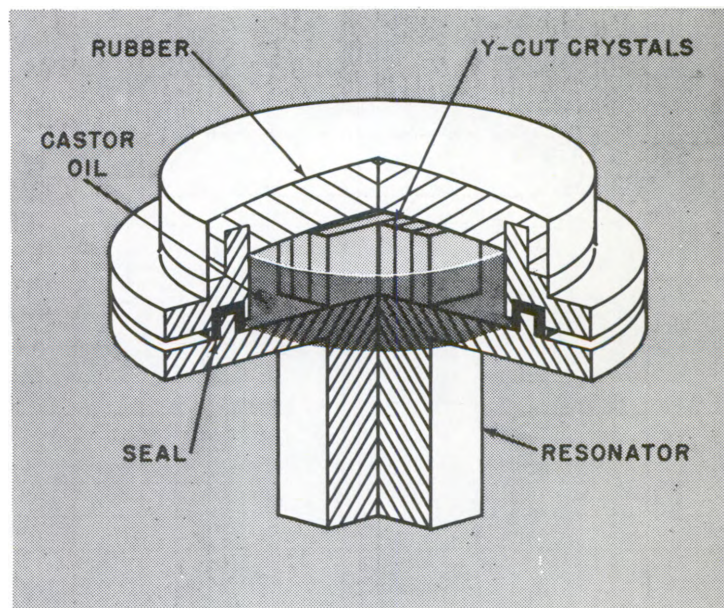
Materials most commonly used for sonar transducers have included quartz, tourmaline, Rochelle salt (sodium potassium tartrate), ammonium dihydrogen phosphate (ADP) and lithium sulphate. Suitable large quartz crystals are difficult to obtain. Most crystal transducers use Rochelle salt and ADP because they exhibit a strong piezoelectric effect and are easily grown to the desired sizes. Because Rochelle salt cannot tolerate high temperatures to which it may be exposed during use or storage (it will disintegrate from internally generated heat if used to transmit high power for any length of time), it has generally been displaced by ADP and lithium sulphate. Because Rochelle salt has the strongest piezoelectric effect, it is still used in listening transducers and for low-power applications. Tourmaline is used for calibration transducers because it is relatively insensitive to temperature changes.

The crystals can be cut and stacked along various crystalline axes, depending on the transducer design. They are cut one-quarter wavelength long for the desired frequency of operation with an orientation such that they will vibrate longitudinally when an a-c field is applied to metal foil glued or deposited on two faces.

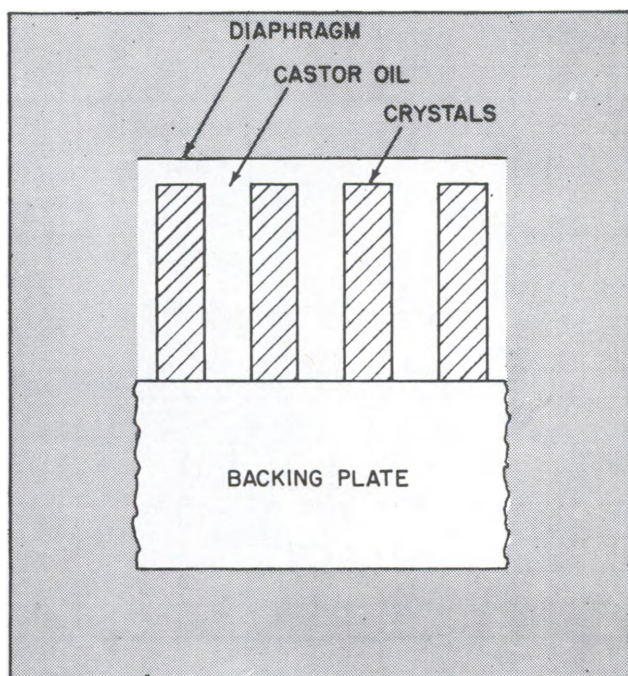
A crystal transducer is constructed by cementing or mounting a stack of quarter-wave crystals to a heavy steel backing resonator plate also one-quarter wave-length thick. The assembly is mechanically resonant at the operating frequency with a node at the interface of the plate and crystals, and maximum amplitude at the free end of the crystals.

Since both Rochelle salt and ADP are soluble in water, the crystal assembly using these materials is a sealed enclosure. To permit sound propagation, the enclosure must be filled with a thoroughly dehydrated liquid. Castor oil is commonly used because the velocity of sound in it is almost the same as that in sea water. Sound propagation from the castor oil to the sea water is via a thin metal diaphragm forming the face of the transducer assembly, shown in the illustration, or through a diaphragm of Rho-C rubber (i.e., rubber which has a "c" acoustic resistance equal to that of the water).

When used for listening, the maximum electric signal is produced by the crystals if the frequency of the received sound is the same as the resonant frequency of the transducer.



ROCHELL SALT CRYSTAL HYDROPHONE



CRYSTAL TRANSDUCER

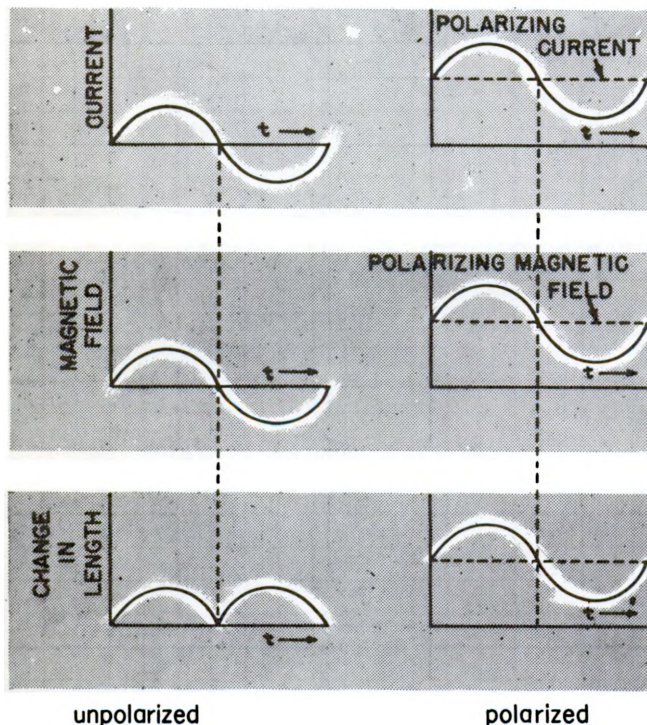
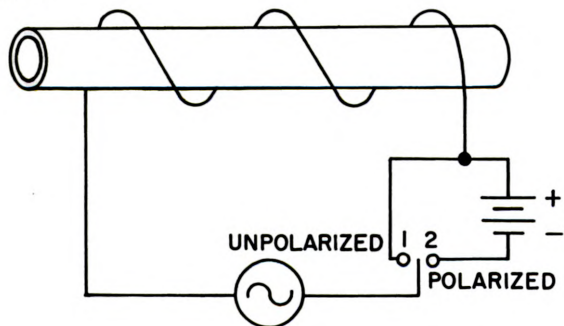
CERAMIC TRANSDUCERS

When an electric field is applied across a dielectric, the dielectric is deformed. This phenomenon of change in dimensions is called electrostriction and is independent of the direction (sign) of the electric field and proportional to the square of the field intensity. In the crystal transducers, the electrostrictive effect is present but much smaller than the piezoelectric effect and so is ignored. However, with barium titanate, a ceramic, the electrostrictive effect is large in comparison with the piezoelectric effect.

Ceramic for transducers has the advantage over crystals that the ceramic can be molded to any desired shape. This property is particularly desirable for making cylindrical scanning or omnidirectional transducers. The converse electrostriction, the change in electric potentials when the material is stressed, takes place only when a constant polarization potential is present. The a-c sonar signal is thus superimposed on the larger d-c polarization and the material dimension will vary directly with the magnitude of the resultant potential. The polarization of electrostrictive materials is thus analogous to the magnetic polarization required for magnetostrictive materials. Without the polarization, mechanical stresses will not produce an electric potential.

magnetostrictive transducers

A magnetic field will cause a number of materials to change dimensions in the direction of the applied field. Depending on the material and the strength of the magnetic field, some materials will expand and some will contract, but independently of the direction of the applied field. Nickel is the most commonly used material in magnetostrictive transducers because it exhibits a very large magnetostrictive effect. Nickel contracts in a magnetic field to an extent largely proportional to the intensity of the field.

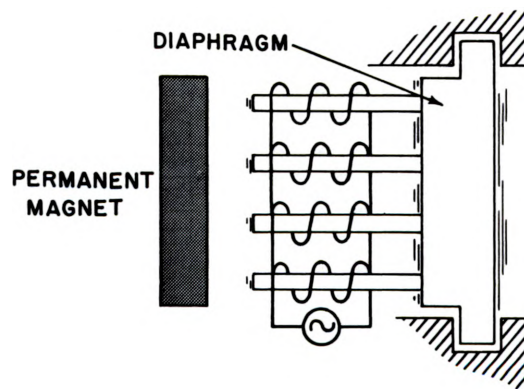
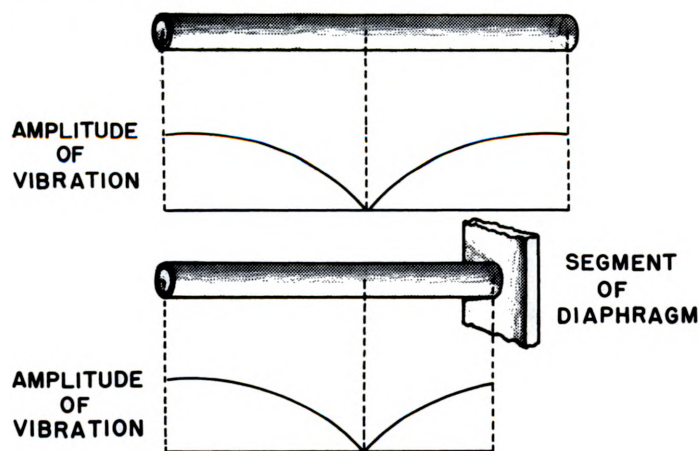


The nickel elements in transducers are generally one-half wave long and are supported at the central node so that the maximum amplitude of vibration takes place at the ends. In one type of transducer the elements are nickel tubes. To increase the area of the active face in contact with the water in this design, a series of tubes or rods are attached to a diaphragm. The diaphragm becomes the active face of the transducer. The dimension of the tubes and diaphragm plate is designed to make the transducer mechanically resonant at the operating frequency for maximum efficiency.

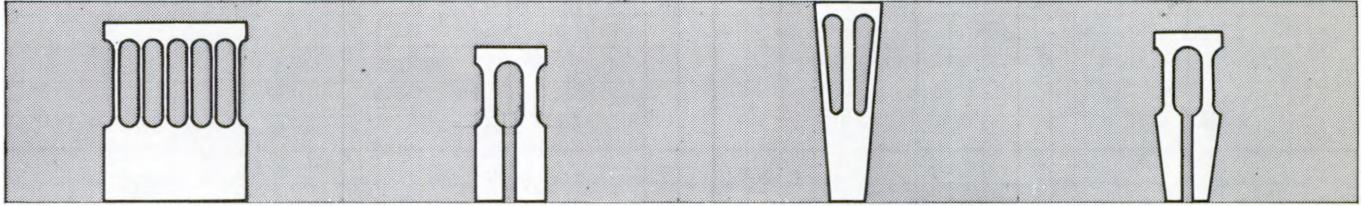
Because the nickel contracts for either direction of magnetic field, an applied a-c sinusoidal field will cause the nickel to contract for each half of the a-c cycle, resulting in a frequency of induced vibration twice that of the applied field. In a transducer, the sinusoidal magnetic field is obtained by winding coils around nickel rods, tubes or stamped elements (stamped and stacked in various shapes), and impressing a current at the desired sonar signal frequency through the coil.

It is generally desired that the vibration and the resulting acoustic signal be of the same frequency as the electric signal applied to the coils. This is accomplished by superimposing a large d-c polarizing magnetic field on the nickel elements by means of a d-c polarizing current through the coil or directly by permanent magnets. If the d-c field is larger than the a-c signal field, then the resulting magnetic field will always be in one direction, with its magnitude varying with the a-c signal. The change in nickel element length will thus be synchronous with the a-c signal.

A d-c polarizing field is also needed for application of the converse magnetostrictive effect. If the nickel element is subjected to varying longitudinal mechanical forces (acoustic pressure waves), the resulting stresses change the magnetic permeability of the nickel. If a d-c magnetic field is present, the permeability changes will cause corresponding changes in the magnetic field intensity. These field changes will in turn induce a varying voltage in the coil wound around the nickel element which will vary proportionally with the impinging acoustic waves. Without the ambient d-c polarizing field, no voltages will be induced in the coils.



The magnetostrictive element can also be constructed of solid stacks of stamped nickel laminations separated by thin plastic insulation (to prevent eddy currents). Openings in the stamped bars are provided for the coils. With this design the active face of the tightly packed elements can be in direct contact with the water. Tapering of the elements simplifies construction of curved or cylindrical transducer faces.



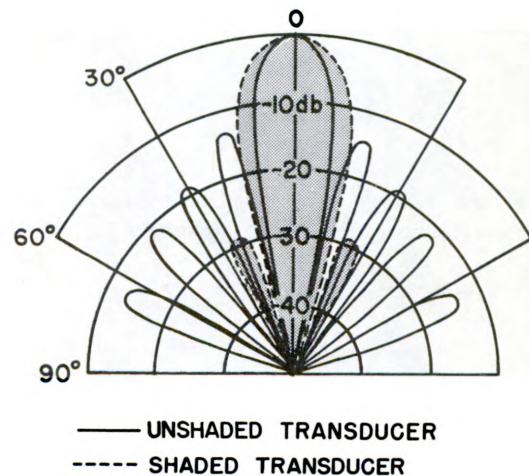
hydrophones

Hydrophones are transducers designed only for reception (listening). They may be either magnetostrictive or electrostrictive and utilize the same principles as for two-way transducers. However, since they do not handle high transmission power (with its resulting increase in temperature) their construction can be much lighter. Also, if a hydrophone is intended for listening over a broad frequency band it is designed with a resonant frequency outside the desired listening frequency range to prevent nonlinear peaking effects.

transducer directivity

The directivity or beam pattern of a transducer is a function of the mechanical arrangement of the transducer elements and the associated electrical and acoustical circuitry. As with antennas, where the directivity pattern is caused by interference between electromagnetic radiation from various parts of the antenna, so in transducers the patterns are caused by interference between sound radiated from various points on the transducer surface. Baffles and shields also serve to establish the desired patterns. Control of transducer patterns is needed for various reasons: the directivity can be matched to operational needs; they can be designed to maximize target bearing determination by using narrow beams for reception and broad beams for transmission; they can be designed to exclude noises from directions, other than along the transducer beam by reducing back and side lobes thus reducing background noises and permitting detection of weaker targets; and they can concentrate power transmitted in one direction to obtain more power on the target for stronger echoes and longer ranges.

Directivity patterns can be modified, such as for the reduction of side lobes by shading a transducer. A transducer is shaded in design if the strength per unit area (power projected per unit area) is varied over the surface. Shading can be accomplished physically by arranging the spacing, size, or number of elements, or electrically by the number of turns on magnetostrictive elements. Directivity patterns also can be varied



electrically by phasing and delay networks associated with the transducer elements and mechanically by the physical shape of the active surface of the transducer. For circular face, flat transducers, the beam width, θ , is given by the equation:

$$\sin \theta/2 = 0.61 \lambda/D$$

where: θ is the angle between half power points
 λ is sonic wavelength
 D is transducer diameter

For satisfactory operational resolution, the beam width must not be greater than 10 degrees. For this width the ratio D/λ must be at least 6. From this it can be seen that low-frequency operation requires large-size transducers.

transducer power and intensity

A transducer transforms electrical into mechanical and acoustic energy by generating an alternating sound pressure that is superimposed on the static ambient pressure of the medium through which the sound is propagating. The relationship between rms acoustic power output P of the transducer and the sound pressure is:

$$P = p^2 A / \rho c$$

where: P = acoustic power in ergs/second
 p = rms pressure of the medium in dyne/cm²
 ρ = medium density in gms/cc
 c = velocity of sound in the medium in cm/sec
 A = area through which the acoustic energy flows in cm².

The product (ρc) is called the acoustic resistance of the medium, and is equal to about 150,000 grams/cm²/sec for sea water. One gm/cm²/sec unit is called the acoustic ohm. Sound intensity I in erg/cm² is defined as:

$$\frac{I}{A} = P/A = p^2 / \rho c \text{ ergs/cm}^2, \text{ or}$$

$$I = (10^{-7} p^2) \rho c \text{ watts/cm}^2$$

At a distance (r) from a point source emitting sound equally in all directions, sound intensity I is:

$$I = (P/4\pi r^2).$$

Combining the above equations yields:

$$P = (10^{-7} P p c / 4\pi r^2) 1/2 \text{ dynes/cms}^2$$

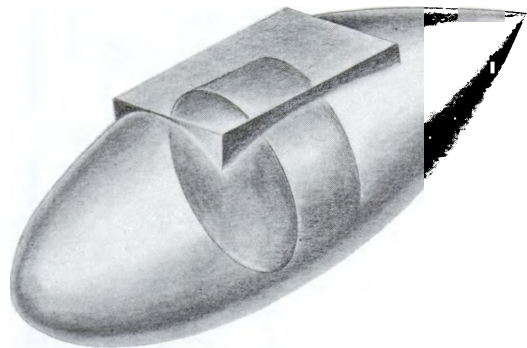
$$P = (10^{-7} \rho 4\pi r^2 p^2) / \rho c \text{ watts.}$$

The power per unit area (P) from a transducer is thus a function of the sound wave pressure, p . The pressure of the sound wave generated at the surface of the transducer is limited by the hydrostatic pressure at the transducer, which in turn is limited by the cavitation pressure, which is the pressure below which water vaporizes. If the pressure in the rarefaction portion of

the sound wave falls below the cavitation pressure, the water vaporizes and small air bubbles form which greatly reduce the transducer efficiency. This factor limits maximum transmitted power to about 2 watts/square inch of the transducer area in the region near the surface of the sea. Maximum transmitted power increases with depth, however, as the square of the ambient pressure. Thus, at a depth where the ambient temperature is doubled, the maximum transmitted power becomes 8 watts/square inch of transducer area.

Increasing transmitted power for a given transducer configuration thus means increasing transducer size. For a given amount of power per unit of surface area, the transducer size is also directly related to signal frequency, since the electro- or magnetostrictive elements must be a quarter- or half-wavelength in size (the lower the frequency, the longer the wavelength), and for a reasonable beam width the transducer face must be 6 to 8 wavelengths wide. Since maximum range (minimum attenuation) is obtained with lower frequencies, long-range sonar entails very large transducers.

This size vs. power and frequency situation presents severe physical and tactical limitations and problems. Large transducers add considerably to drag effects (even with streamlined domes). They present problems on how and where to mount and drive them. They require very large power plants. They require large ships to accommodate them and their associated electrical and electronic equipment. Some of the new long-range sonars are limited, therefore, to installation in cruisers or larger ships.



New methods of exciting water with high-power sonic waves are required. Various explosive pneumatic and hydraulic sound sources are under development for this reason.

The total amount of power transmitted by a transducer of fixed size and frequency is limited at the moment, and therefore concentrating the power in a beam by control of directivity patterns becomes the more practical method of increasing power to a target.

typical searchlight system

In searchlight echo-ranging systems, a short burst or pulse of sound energy (called a ping) is transmitted from a highly directive transducer through the water. Range is measured by accurately determining the time interval between the transmitted burst and the instant its reflection returns from a target or discontinuity in the water (ship, submarine, wake, etc.). Bearing is determined by noting the direction in which the transducer is trained when the intensity of the reflected signal is greatest. The term searchlight applied to this type of equipment comes from a comparison of the sharp beam projected by the transducer with the beam of a searchlight.

The searchlight transducer is generally a flat-faced transducer designed and shaped to produce a very sharp forward directivity pattern. In some cases a switch arrangement with the transducer elements permits a broad transmission pattern and a narrow reception pattern. The transducer must then be physically rotated or trained in order to scan an area or to determine the bearing of a target for which the intensity of a reflected signal is greatest.

range measurement

Various types of range indicators are used with searchlight echo-ranging equipment. In one case, a disk rotated at constant speed is the timing element. This disk has a radial slot near the edge with a light mounted behind it. The equipment is keyed and a pulse of sound energy is transmitted when this slot goes past a zero position. An echo from a target causes the light behind the slot to flash. If we let

c = velocity of sound in the sea

R = range

Δt = the time interval between the instant a pulse of sound energy is transmitted and the instant an echo is received from a target

ω = angular velocity of the disk

a = the angle between the zero position of the radial slot in the disk and the position of the slot when a returning echo causes the light to flash

then

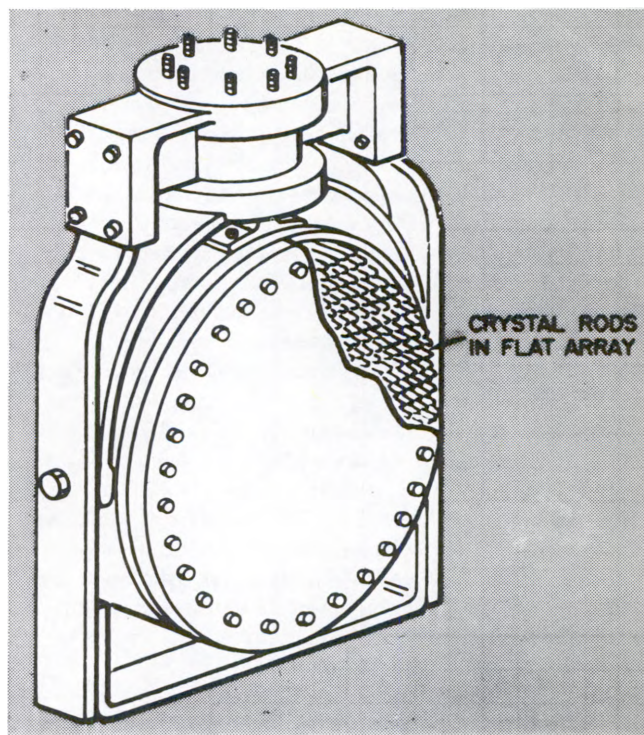
$$R = \frac{c\Delta t}{2} \text{ and } a = \omega\Delta t.$$

Dividing the two equations yields:

$$R/a = c/2\omega.$$

Solving for range:

$$R = \frac{c}{2\omega} a.$$



From this range equation it is seen that the angular position of the radial slot at the instant a reflected signal causes the light to flash is a measure of range. A ring around the disk is calibrated in range so that range can be read directly against the position of the flash.

Another type of range indicator utilizes a recorder in which a stylus drawn at constant speed across a roll of treated paper is the timing element. A target echo causes a voltage to be applied between the stylus and a metal plate under the paper, marking the paper. The distance from the zero position of the stylus at one edge of the paper to this mark is proportional to range, which can be read off from a calibrated scale.

The paper, carried in a roll, is slowly drawn under the stylus at right angles to the stylus motion. Thus a new surface is presented for each trip of the stylus and a permanent record is made of the target echoes received.

The chief advantage of this type of range indicator is that information about the target can be gained from the pattern formed by the traces from successive rangings that is not readily available from the flashing lamp indicator.

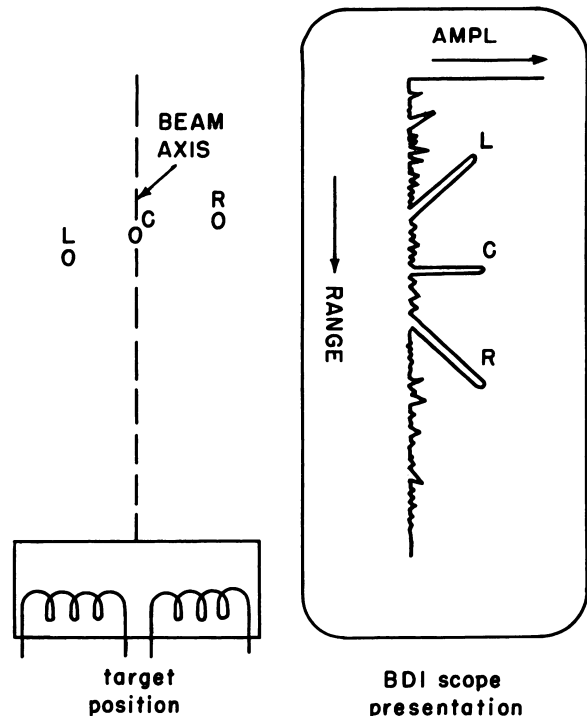
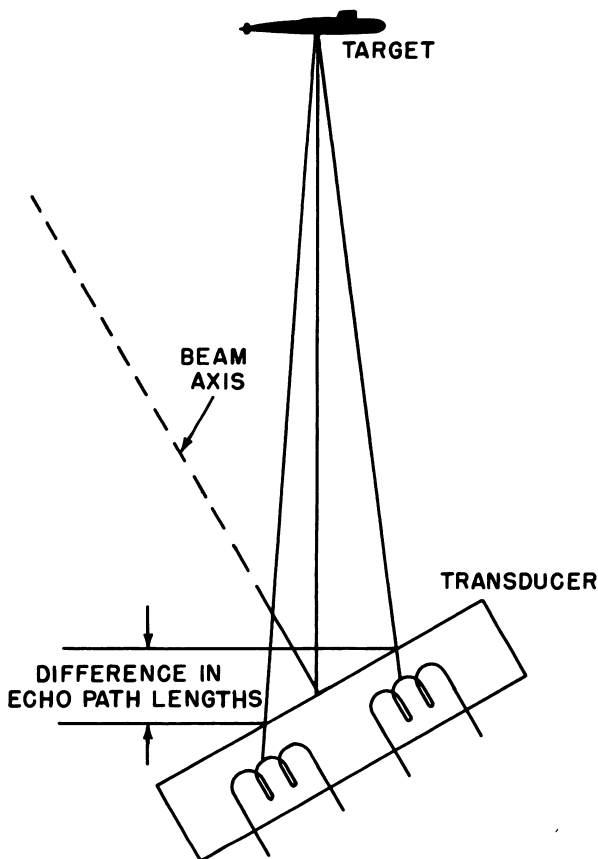
A third type of range indicator utilizes a cathode-ray tube similar to a standard radar range scope. A horizontal sweep is initiated upon transmission and the return echo is seen as a vertical pulse or pip on the cathode-ray tube screen.

bearing measurement

Bearing of the target is determined by observing the direction in which the transducer is trained when a maximum signal is received. If this direction is determined solely by listening to the amplitudes of successive echoes or observing a device which indicates the amplitude of the echoes, the method becomes slow and inaccurate. The amplitudes of successive echoes are influenced by fluctuation as well as by the position of the target in the transducer beam; in addition, a succession of observations is necessary to determine a bearing, thus limiting the rate at which bearings may be determined on a moving target and permitting a change in target bearing while the observations are being made. To reduce these difficulties, target bearing indicator devices were developed. Such devices indicate on which side of the axis of the transducer beam a target lies if the beam is not directly on the target, and give a rough indication of the angular distance from the beam axis. If the target is not on the axis of the transducer beam, there is a difference in path length for the left and right halves of the transducer. The two halves are connected separately and the electrical output is brought out separately. The path length causes a phase difference between the two outputs, the output from the left lagging in the case shown with respect to the output from the target to the right of the beam axis. The outputs of the two halves of the transducer are amplified and fed into phase-comparison circuits which produce a signal which has an amplitude proportional

to the phase difference and is positive or negative, depending on which half has the lagging output. The output of the phase-comparison circuits is placed across the horizontal deflection plates of a cathode-ray tube (crt) in the bearing deviation indication (BDI). A voltage which increases linearly with time, starting from the instant when the equipment is keyed, is also placed on the horizontal deflection plates to produce a deflection increasing to the right with time. A sum signal is placed across the vertical deflection plates.

With no received signals, the trace on the crt in the BDI is a straight horizontal line. An echo from a target not on the axis will cause the phase comparison circuits to produce an output causing a deflection of the trace up and to the right when the target is to the right of the beam axis, and up and to the left when the target is to the left of the beam axis, as shown in the illustration. Because of the sweep voltage applied to the horizontal deflection plates, this deflection will occur at a distance from the beginning of the trace proportional to the range of the target. A target on the beam axis causes an upward deflection of the beam. The angle of the deflection or leaning (hence the name Pisa indicator is sometimes used) gives a rough indication of the amount that the transducer will have to be trained to bring the beam axis on the target. The signals from the phase comparison circuits can also be fed to serve circuits for automatic training of the transducer in automatic tracking operations where the sonar is used for fire control purposes.



depth measurement

Two forms of depth measurements are involved in sonar echo-ranging systems: bottom sounding, and target depth determination. In bottom sounding, the same techniques for range measurements are used to measure sea bottom depth, except that a downward-facing transducer is required.

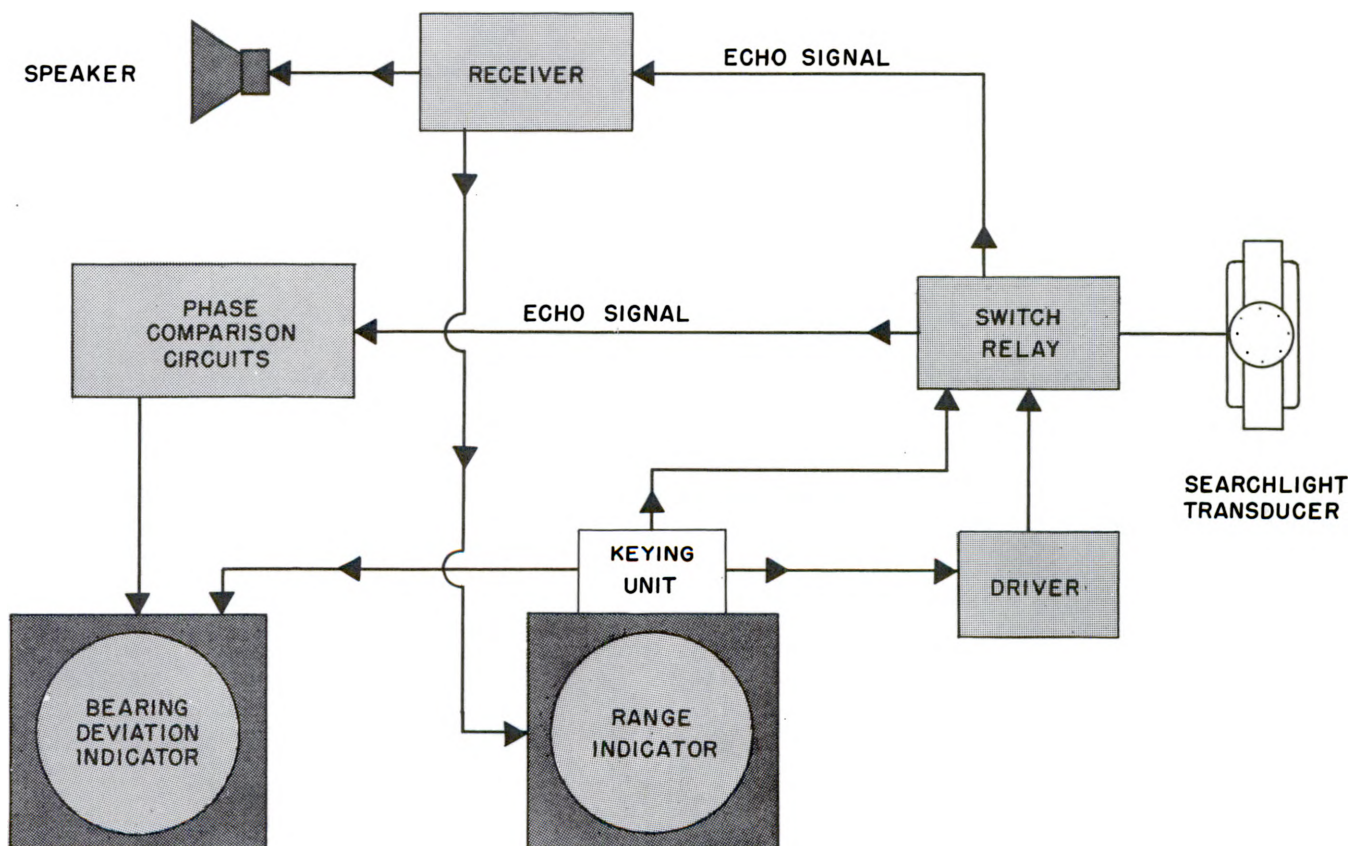
Target depth measurement is more complex. One method is to use a technique similar to that for bearing measurement. A transducer is required that has upper and lower elements as well as left and right elements and that can be moved in elevation as well as bearing.

A similar set of phase-comparison circuits and crt indicator as used for the BDI will determine the elevation angle when the center of the sound beam is on the target. The transducer will then be pointing at and give the elevation angle to the apparent target. Since the sound path from the real target will have been refracted and bent by varying oceanographic conditions before it reached the transducer, various correction factors must be computed to ascertain the true target depth below the transducer. The depth of the transducer below the water will also have to be taken into account if true target depth below the surface is desired.

functional operation

A cycle of operation is initiated with a signal from the keying unit. This unit is usually associated with the range indicator and produces its signals when the range indicator is in the zero range position. Signals are sent to three units to perform the functions indicated. They are sent to the switching relay, to cause this relay to disconnect the transducer from the receiving circuits and connect it to the driver. At the end of a

predetermined length of time (a little longer than the transmitted pulse) this relay reconnects the transducer to the receiving circuits (to prevent damage to the receiver). They are sent also to the driver, to trigger the driver, and to the BDI, to initiate a horizontal sweep. When triggered by a signal from the keying unit, the driver transmits an electrical pulse of the proper amplitude and frequency through the switching relay to the transducer. The transducer converts this pulse of electrical energy to sound and radiates it in a beam.



Any objects in the sound beam will reflect back to the transducer a part of the sound energy falling on them. The transducer transforms this reflected sound energy into electrical energy. This electrical signal goes from the transducer through the switching relay (which returned to the receiving position shortly after the end of the transmitted pulse), to the receiver, and to the phase-comparison circuits of the BDI channel.

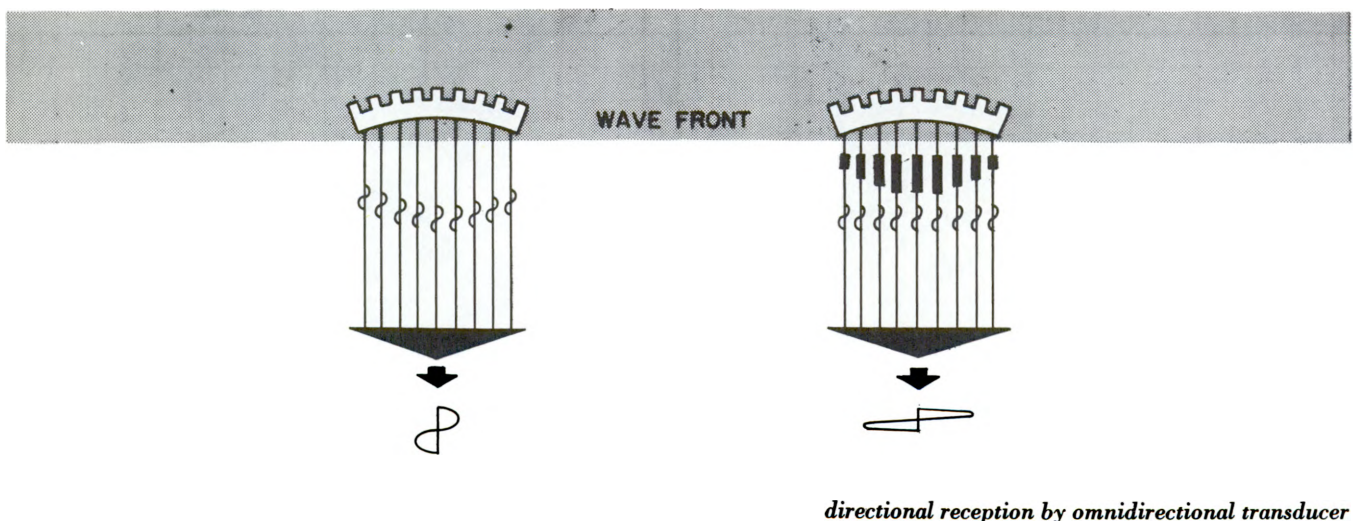
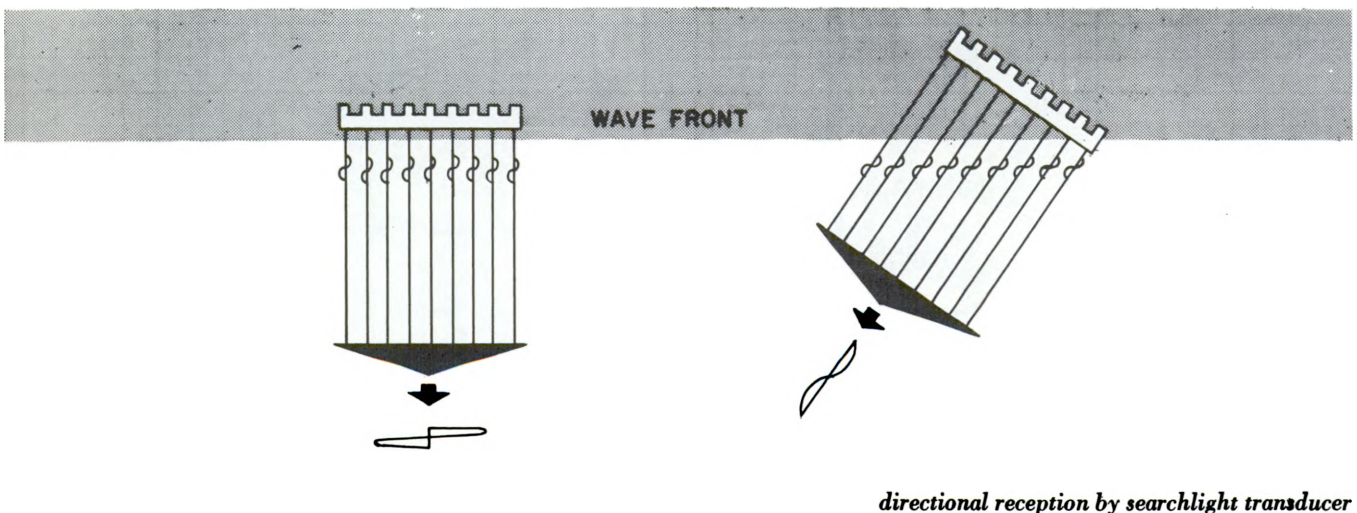
In the receiver, the received signals are amplified and combined with the output of a local oscillator to give a beat frequency in the audible range of frequencies. The audible output is fed into a loudspeaker or earphones. A part of the received signal, after amplification, is rectified to give a d-c signal proportional in amplitude to the received signal. This output from the receiver goes to the range indicator.

At the end of a period of time dependent on the distance to which search or ranging is being conducted, the keyer produces another signal and the cycle is repeated. If search is being conducted to 5000 yards, then the pulse

repetition frequency (prf) must allow an interval long enough for sound to travel to a target 5000 yards away and back to the transducer at an average velocity of 4800 feet per second.

For shorter search ranges or when ranging on a target closer than 5000 yards, the interval may be shortened by increasing the prf. Provisions are made in the keying unit for varying the prf.

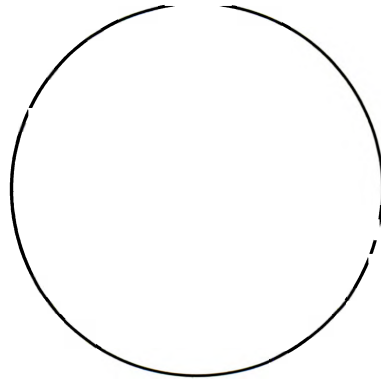
Any sound arriving from a direction in the transducer beam and with a frequency within the frequency range of the transducer, even though not a reflected portion of a transmitted pulse, will be received. Such ambient sounds can be heard in the speaker and produce indications on the range indicator and BDI. As such sounds bear no fixed time relation to the transmitted pulse, the range indications are meaningless. A continuous sound will cause a range indication. However, frequently it is possible to identify the source of such sounds by their characteristics and determine the bearing of the source. Sonar equipment may be operated without transmitting to obtain the information that such listening can provide.



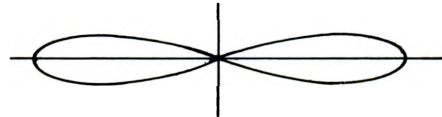
typical scanning system

The principles used in the scanning echo-ranging equipment for the determination of range and bearing are the same as those used in searchlight equipments. In both, range is determined by accurate measurement of the time interval between the instant a burst of sound energy is transmitted and the instant a reflected signal is received. In both, bearing is determined by noting the position of the transducer beam when the intensity of the reflected signal is greatest. The two types differ greatly in the method of presenting range and bearing information, the type transducer used, the method used to sweep the transducer beam and the rate at which the transducer beam is swept.

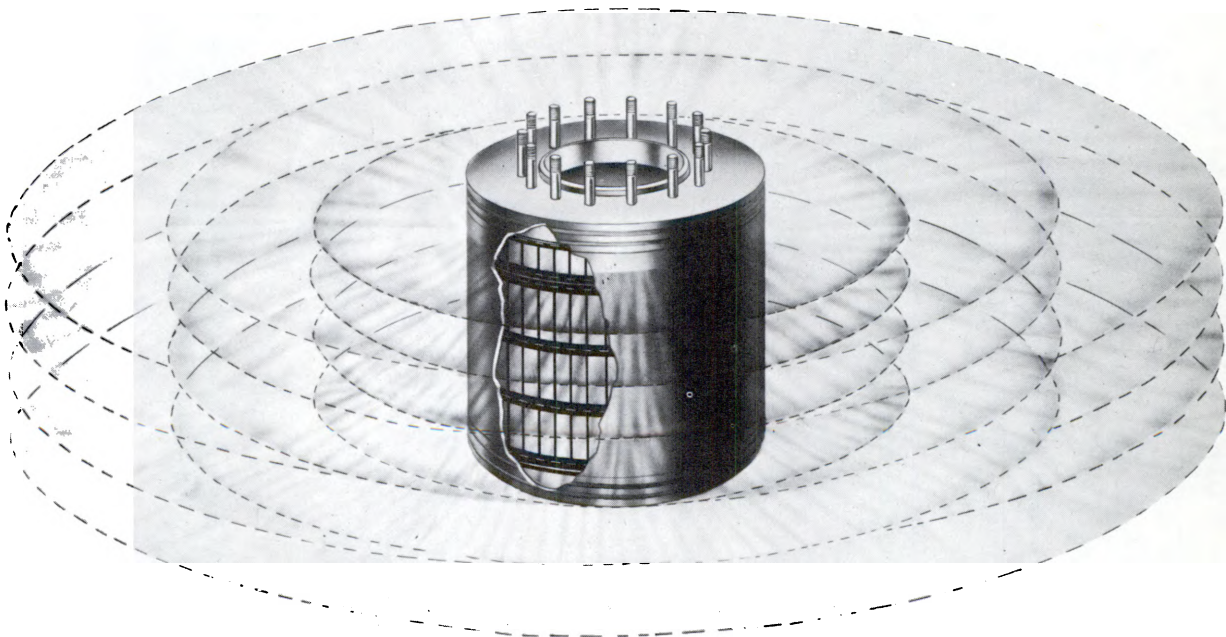
In sonar scanning equipments, a fixed cylindrical transducer is used. When transmitting, sound energy radiates from the transducer with equal intensity in all directions. When receiving, the scanning switch, acting in conjunction with the beam-forming network, causes the transducer to be sensitive only to signals returning from a narrow sector called a beam. This beam is rotated rapidly by means of the scanning switch described in the following paragraph and the range and bearing information is presented on a cathode-ray tube as a polar or PPI plot. The use of a rapidly rotating beam makes it possible to obtain range and bearing on every target in range of the equipment for every transmission.



PATTERN IN HORIZONTAL PLANE



PATTERN IN VERTICAL PLANE

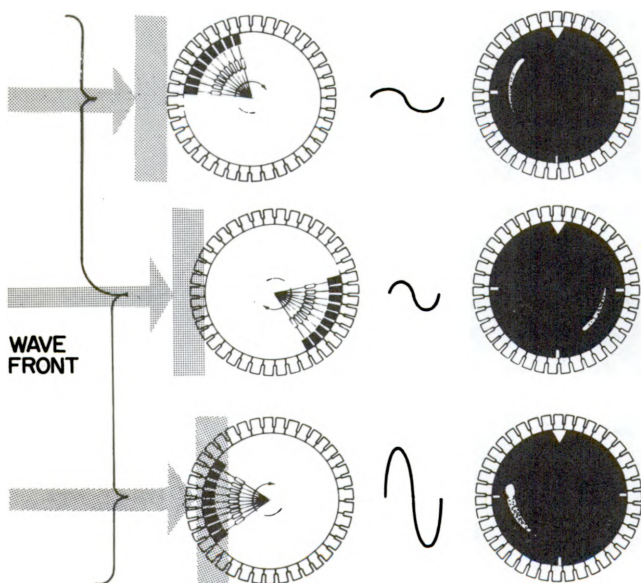


scanning switch

One of the principal advantages of a scanning search transducer over a searchlight transducer is that it can scan electronically by means of a scanning switch while a searchlight transducer must scan mechanically. To scan the entire area around the ship, the delay circuits which compensate for the time delays on the individual staves of the transducer due to its cylindrical shape, are connected to rotating contacts instead of directly to the transducer stave. These rotating contacts comprise the scanning switch. The position of the scanning switch at which a maximum signal is received indicates true target bearing (provided the transducer is oriented in the horizontal plane). This information can be displayed on a crt which has a light trace rotating at the same frequency and in phase with the scanning switch. The point at which the rotating light trace exhibits maximum intensity is the bearing of the target.

To receive echoes regardless of their directions and times of reception, the scanning switch must rotate at least once in each pulse length.

A typical scanning switch contains a rigidly mounted glass disk about 11 inches in diameter. One side of this disk is silvered and then scribed with 48 radial lines to form 48 separate conducting sectors, each shaped like a slice of pie. Each segment is connected electrically through a hole bored in the glass disk to a metal pin on the back of the disk. One of the 48 staves of the transducer is electrically connected to each of these pins. A second glass disk similar to the first is mounted on a shaft with its silvered surface a few thousandths of an inch from the silvered surface of the first. About 16 adjacent sectors (the number varies with the equipment) of this second disk are connected to the beam-forming network, which is mounted on the shaft. The output of the beam-forming network is taken off by slip rings and led out of the scanning switch.



When the two glass disks are positioned so that their conducting sectors are aligned, they form small capacitors. The beam-forming network delay lines are connected electrically to 16 adjacent staves of the transducer through 16 of these capacitors. The network combines the outputs of the 16 staves to give an output equivalent to that of a linear array.

When the shaft-mounted disk is rotated, the 16 segments to which the network is connected rotate around the circumference of the stationary disk, and successive groups of staves are connected to the network to produce a rotating beam.

A typical scanning sonar equipment contains two of these scanning switches: one for the video channel and one for the audio channel. The video channel furnishes the signals for the PPI presentation. This scanning switch for the channel typically is driven at a constant speed of about 1800 rpm, giving about 30 scans per second. The audio channel output is fed into a loudspeaker to permit aural evaluation of target echoes. The audio scanning switch does not rotate continuously, but is positioned by a servo system controlled by a handwheel at the indicator unit.

ACTIVE AREA DURING SCANNING

At any time after the transmitted pulse, in a scanning sonar, it is seen that the target reflecting the transmitted pulse and producing a response in the transducer must lie in an area in the horizontal plane bounded by the beam width and half the pulse length. This area, called the active area, forms an expanding spiral as the beam rotates. The range, R , to the center of the active area is given by the equation for the spiral:

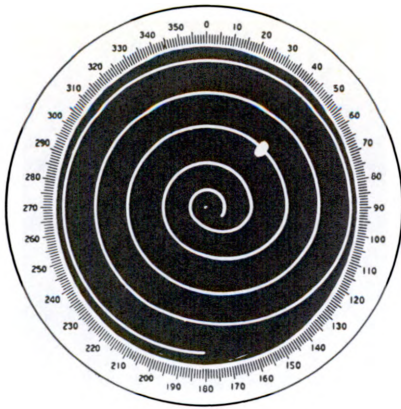
$$R = \theta c / 2\omega$$

where θ is the direction of the beam and ω is the angular velocity of the rotating beam.

To insure that any target in the scanned area covered between transmitted pulses is detected, the pulse length must be equal to or greater than $2\pi/\omega$.



path of active area for scanning sonar



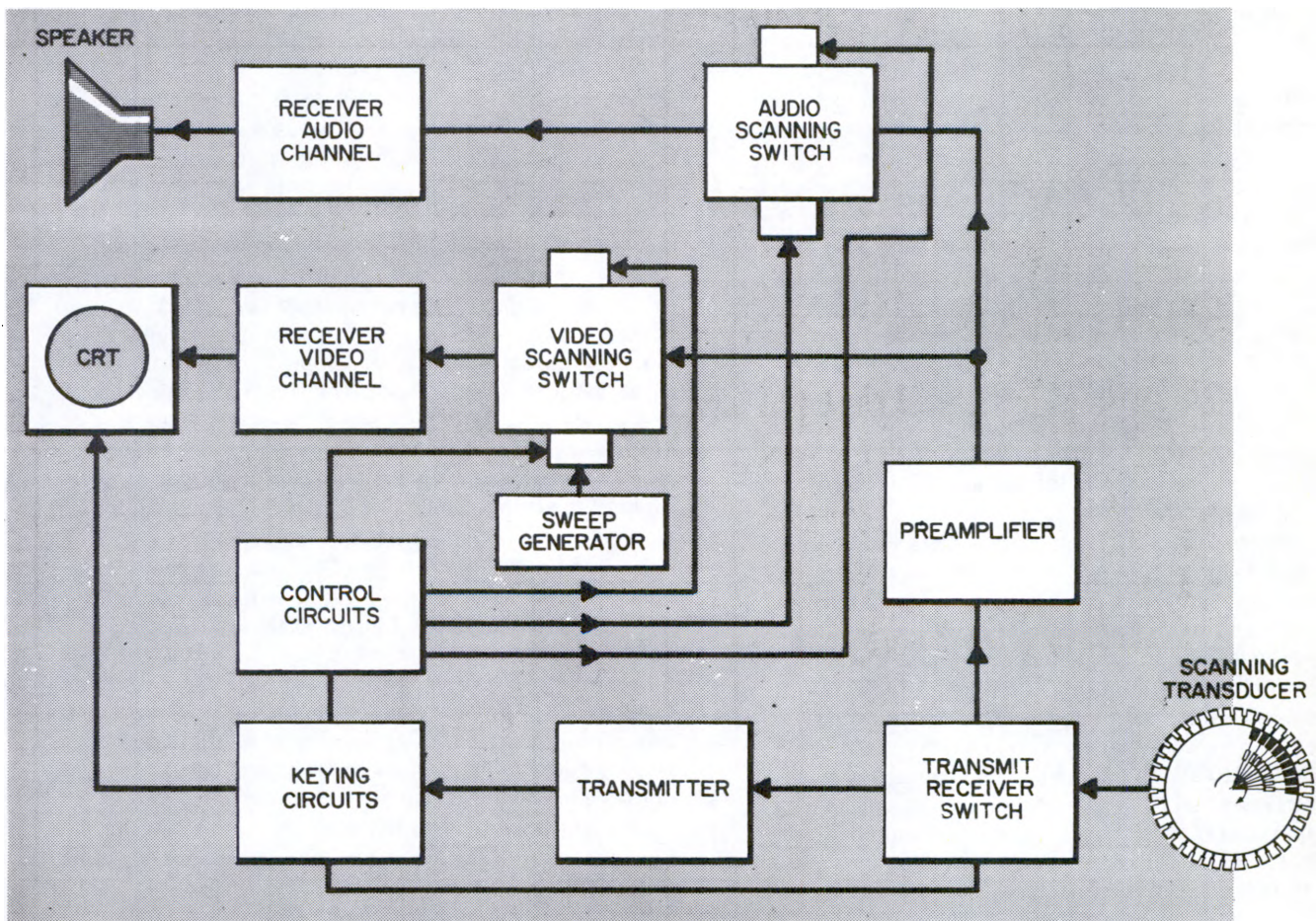
cathode ray tube display

For presentation of range and bearing information, the scanning echo-ranging equipments use a PPI presentation similar to that used in radar. To obtain such a presentation, the electron beam in a cathode-ray tube is deflected so that the trace on the face of the tube follows a spiral path similar to that of the active area shown in the illustration. A voltage is applied to the grid of the tube to produce an intensification of the trace when a target echo is being received, resulting in a presentation on the tube that is a scale plot of the targets in the surrounding area.

functional operation

A functional block diagram of a typical scanning sonar equipment is shown. A cycle is initiated by signal pulses from the keying circuits. These signal pulses go to the following four units, where they perform the functions indicated.

- 1) Transmit-receive switch. Switches the transducer from the receiver circuits to the transmitter circuits. The transducer is reconnected to the receiver circuits at the end of a time interval of about 100 milliseconds.
- 2) Transmitter. Triggers the transmitter and causes the formation of a pulse of the proper length, frequency and amplitude.
- 3) Sweep generator. Returns the sweep generator to zero range conditions and starts a new cycle.
- 4) Control circuits. Causes control circuits to disconnect the cathode-ray tube deflection coils from the synchro generator on the video scanning switch and connect it to the one on the audio scanning switch. Also causes the control circuits to generate a positive blanking voltage which is applied to the cathode of the cathode-ray tube to prevent spurious traces. This blanking voltage is removed for a short period during the keying interval to permit formation of the audio strobe trace. At the end of the keying interval, the control circuits return to the receive condition.



The pulse formed by the transmitter goes through the transmit-receive switching section and is applied to all staves of the transducer with the same phasing. The transducer therefore transmits a burst of sound of the same duration and frequency as the exciting pulse, dispersed equally in all directions.

Target echoes or independently generated sounds of the proper frequency reaching the transducer cause electrical signals to be generated in the staves of the transducer. These signals go through the transmit-receive switching section, now in the listening condition, to the preamplifiers. In the preamplifiers, one for each staff, the signals are amplified and fed to the scanning switches. The video scanning switch, rotating at about 1800 rpm, takes these inputs and combines them to give an output equivalent to that of a highly directional, rapidly rotating, flat transducer. The output of the video scanning switch is further amplified in the video channel of the receiver and is rectified and coupled to the grid of the cathode-

ray tube in the indicator to produce a brightening of the trace when sound signals are being received.

Signals from the preamplifiers enter the audio scanning switch, where they are combined to give a signal equivalent to that from a highly directional transducer trained on a bearing determined by the position of this switch. The audio scanning switch is positioned by a training signal from the control circuits to the audio scanning switch drive motor. The output of the audio scanning switch is amplified and converted to the desired audible frequency range in the receiver audio channel. The output of the receiver audio channel goes to a loudspeaker to permit aural interpretation of received signals.

When the voltage produced by the sweep generator (the voltage which controls the rotor current in the control transformer) reaches the value corresponding to full deflection on the crt, it triggers the keying circuits, initiating another cycle.

comparison of searchlight and scanning systems

The searchlight sonar has certain limitations that increasingly restricted its usefulness as the underwater speed and range of submarines increased. The chief limitations are listed below.

Slow Search. After transmitting a ping, it is necessary to remain trained on the same or nearby bearing for a length of time equal to that required for sound to travel to a target at the greatest range to which search is being conducted and to return. For a search range of 5000 yards, this interval is nearly 6 seconds. If search is conducted at 5-degree bearing intervals (advisable for good coverage), 3 minutes will be required to search through a 180-degree sector centered on the bow of the ship. In this time a ship or a submarine traveling at 20 knots will travel 2000 yards. If search range is increased, the time to search a given sector increases proportionately.

Easily Saturated. When the equipment is being used to determine range and bearings on one target, it cannot be used to range on another target on a different bearing nor can it be used for search.

Necessary to Train the Transducer. This requirement is important partly because of the generally heavy and cumbersome training equipment which it makes necessary. Of equal or greater importance is the fact that this requirement limits transducers to comparatively small sizes. This limitation on the size of the transducer has two important effects. One, it limits power output. As pointed out previously, cavitation limits power output to about 2 watts per square inch of active transducer surface. Two, it makes necessary the use of comparatively high frequencies. As mentioned previously, the diameter of the transducer face must be 6 to 8 wavelengths to achieve the necessary beamwidth. It follows that if the physical size of the transducer is

limited, there is a lower limit on the frequencies that can be used. Because of this limitation, searchlight sonar frequencies fall in the range from 20 kc to 30 kc. The absorption for these frequencies is high. The effective range of these sonars is limited under most conditions to 1500 to 2000 yards.

To overcome some of the limitations of searchlight sonar, the scanning sonar was developed. The scanning principle permits 360-degree search on every transmission. Each transmission permits range and bearing information to be obtained on a number of targets. The audio channel may be used to examine targets of particular interest more closely. If found necessary or desirable, further audio channels can be added. In scanning sonar systems, the transducer is fixed. This simplifies the mechanical problems encountered in the sonar set and permits the use of larger transducers. The use of greater power is thus possible because of the greater area of the transducer. This advantage is offset to some extent by the fact that the power is propagated in all directions rather than concentrated in a beam, as it is in the searchlight sonar, with a consequent loss of effective range. The scanning transducer is especially useful at lower frequencies because the much lower attenuation of sound at these frequencies makes search possible at much longer ranges.

In spite of the fact that the development of the scanning sonar represented a great advance and eliminated or greatly reduced many of the most important limitations of previous echo-ranging systems, there are still strict limitations on the capabilities of scanning radar, largely imposed by the characteristics of propagation of sound in water.

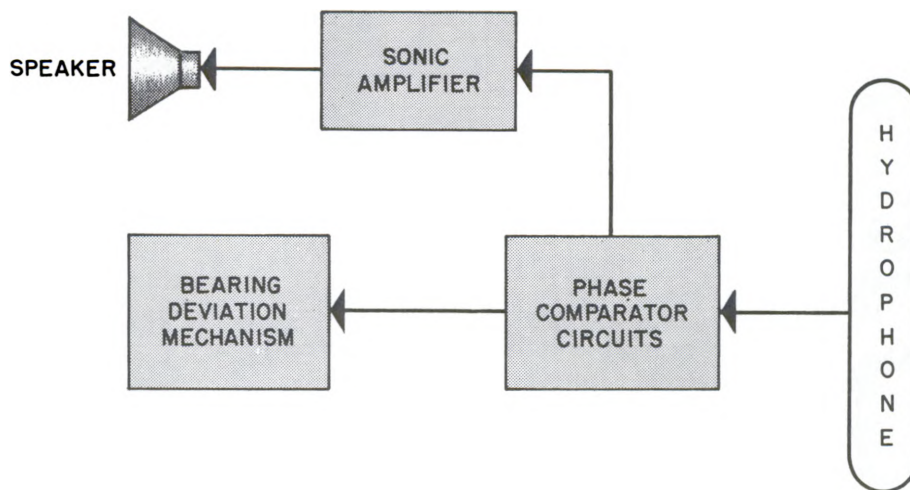
typical listening system

The propellers, machinery and varied activities aboard ships and submarines produce sound that is transmitted to the water, either directly from the turbulence surrounding propellers or through the ship's structure in contact with the water. These sounds can be detected and identified at considerable distances with suitable equipment. The search for, detection and identification of noises produced by ships and submarines is termed listening.

The receiving channels of echo-ranging equipments can be used for listening. Target bearings are determined in the same manner as for targets detected by echo ranging. However, these equipments are needlessly complex for installation where the equipment is to be used solely for listening. Moreover, the frequency range over which echo-ranging equipments have a good response is not the optimum for listening.

An equipment intended to be used for listening is in principle the same as the receiving channel of an echo-ranging equipment, as may be seen by comparing the

functional block diagram of a listening equipment with the functional block diagram of the searchlight echo-ranging system that was shown previously. However, no range indicator is provided, as range cannot be conveniently determined from sounds originating from a target. Another major difference is in the transducer, called a hydrophone when used for listening exclusively. The construction of hydrophones can be much lighter than that of transducers because they do not have to handle high power. They are generally designed with a flat response over a much wider frequency range than transducers. This frequency range is generally below that of transducers, partly because sound is attenuated less at lower frequencies and partly because of the high intensity sounds produced by machinery and propellers in the lower frequency ranges. Ships are frequently detected at long ranges using these equipments. For maximum range sensitivity and bearing resolution some listening installations use large arrays of hydrophones.



TACTICAL APPLICATIONS

Sonar equipment mounted to the hulls of ships presents various problems. Self-generated noises due to water action around the sonar dome require that extreme care be exercised in the design of the dome to provide adequate streamlining. Streamlining is also required to minimize drag. The drag of the sonar dome, particularly for large, long-range sonar, can be considerable, so that in some applications retractable domes are used. Shielding is required to block off own ship's propeller noise, and vibration mounts may be needed to isolate the transducer from shipboard vibration. In one application, the problem of the interference of own ship's hull noise with sonar operation has led to placing the transducer in a streamlined dome in the bow of the ship; this application has the disadvantage that the bow dome is easily damaged and requires careful maneuvering of the ship during docking.

variable depth sonar

To overcome the difficulty of hunting or watching for submarines which take advantage of thermocline shadows for hiding, towed variable depth sonar (VDS) is used. In this application a streamlined fish containing sonar equipment is towed behind a ship, connected electrically to display and control equipment on shipboard. Controllable vanes and depth sensors permit the fish to be towed at any desired depth under the thermoclines.

The use of a towed fish is also an excellent means of isolating the sonar from own ship's noise. Although range data from the VDS fish is adequate, bearing data is not too accurate because of the yawing and pitching of the fish, which is difficult to prevent.

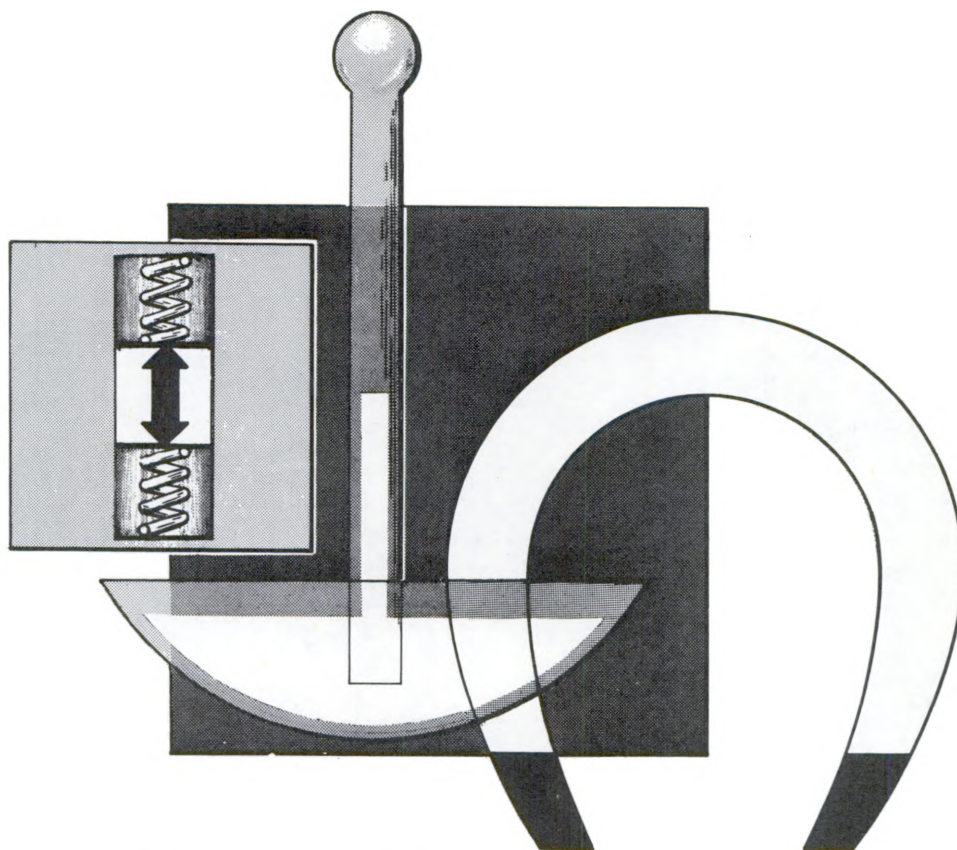
Sonar fish are also towed by helicopter in a form called dipped sonar. In this application, the fish generally houses only the transducer.

Sonar is used on submarines in both active and passive forms. On sub-hunting subs, however, passive sonar is of particular importance. Such subs must operate passively in absolute silence for self protection. These subs generally have very large listening arrays housing multiple hydrophone elements to increase sensitivity and obtain better correlation of received noise signals for maximum bearing accuracy and target identification.

fixed sonar

Fixed sonar installations, both active and passive, are used for harbor and coastal defense. Very large and sensitive pieces of sonar equipment have been placed on the ocean bed along the continental shelf, along the coasts, and near the entrance to major harbors. These installations are connected by cable to shore-based equipment or to radio buoys for transmission of data.

Weapons of various sorts are equipped with sonar as sensing devices, including torpedoes, depth charges, mines, and antisubmarine missiles. Torpedoes and depth charges have been equipped with both passive and active sonar, depending on application and intended target. Sonar-equipped mines are of the passive type to prevent detection. Torpedoes and missiles use sonar equipment as sensors for homing devices. Depth charges, mines, and some missiles use sonar as a proximity detonating device.

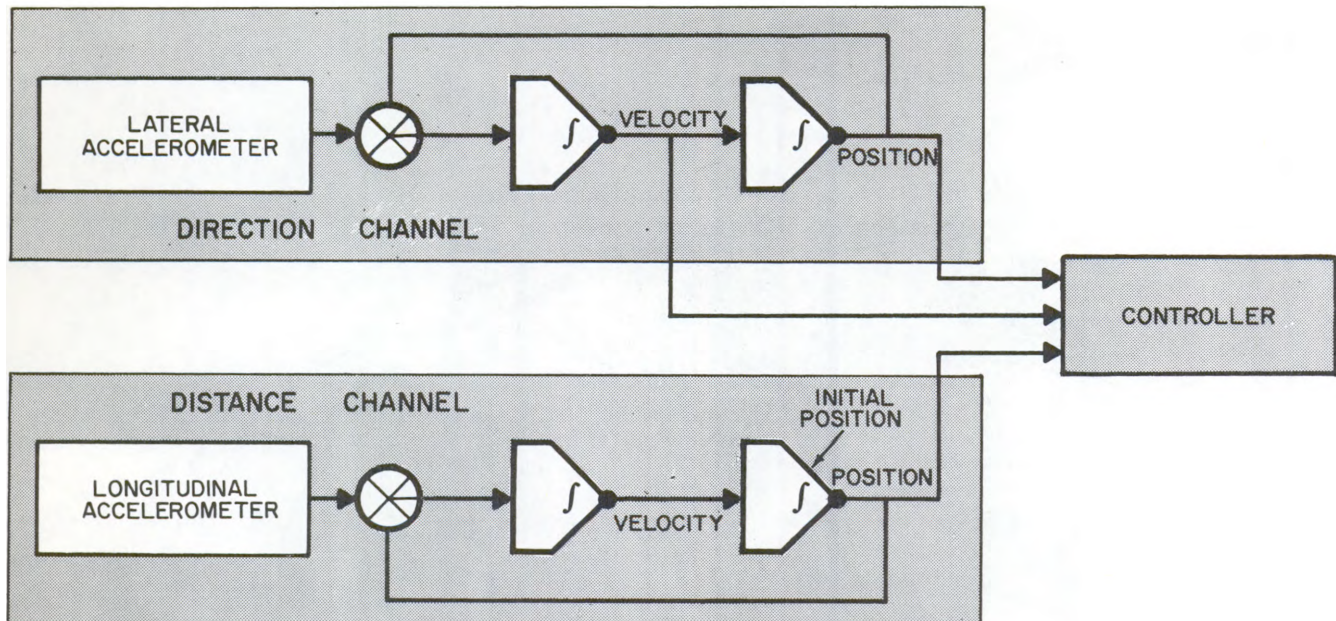


MECHANICAL and MAGNETIC SENSORS

Mechanical and magnetic sensors are other types of devices used for the collection of positional or locational target data. Mechanical sensors include both inertial and pressure sensors (aerodynamic and hydrostatic) that are commonly employed for use in data gathering elements of weapon control systems or as integral components of missile guidance systems.

Magnetic sensors measure variation in the earth's magnetic field which result from the presence of magnetic-field-attracting target structures located within the field. These systems are limited in use to regions that are not located near the earth's poles. Magnetic systems are often employed as auxiliary equipment to various other systems.

MECHANICAL and MAGNETIC SENSORS



inertial sensors

Inertial sensors are used primarily in the inertial guidance or self-contained dead reckoning systems of various missiles and aircraft. The basic principles of inertial guidance are simple. Relative to a known reference frame, sensed and maintained by gyros, the acceleration of the missile is measured in three coordinates by accelerometers. The measured acceleration is integrated once by a computer to obtain missile velocity, and then integrated a second time to obtain missile position. The values of velocity and position are compared with programmed values describing the desired flight path and the resulting error signals, if any, are used to make necessary corrections.

The reference frame is a gimbal-mounted platform stabilized by gyros. (The principle of operation of gyros has been described adequately in Volume 1.) On the platform are mounted three mutually perpendicular accelerometers. Many kinds of accelerometers have been used. All are based on the physical law that:

$$F = Ma.$$

A known mass, M , will require a force, F , to give it an acceleration, a . The basic principle of operation of an accelerometer consists of the measurement of the inertial reaction force of a mass to an acceleration.

TYPES OF ACCELEROMETERS

There are two common types of accelerometers. In the first type the inertial reaction force of the mass causes a displacement of the mass in an elastic mounting system. The displacement, which is proportional to the force and thus to the acceleration, is then measured by

any of several methods. A damper is generally added to reduce resulting oscillation.

In the second type, the inertial deflection of the mass, when accelerated, is detected and a force is instantly applied to the mass to prevent further motion. Acceleration is indicated by the magnitude of the means used to produce the balancing force applied to the mass. These means can be the electric current in a magnet or the pressure of a stream of air used to apply the restoring force.

pressure-sensing devices

There are many applications in which the pressure of a medium must be sensed. All the pressure-sensing devices operate on the basic principle that in any medium pressure is the result of the impact of the molecules of the medium on the device. The density of the molecules is a function of pressure.

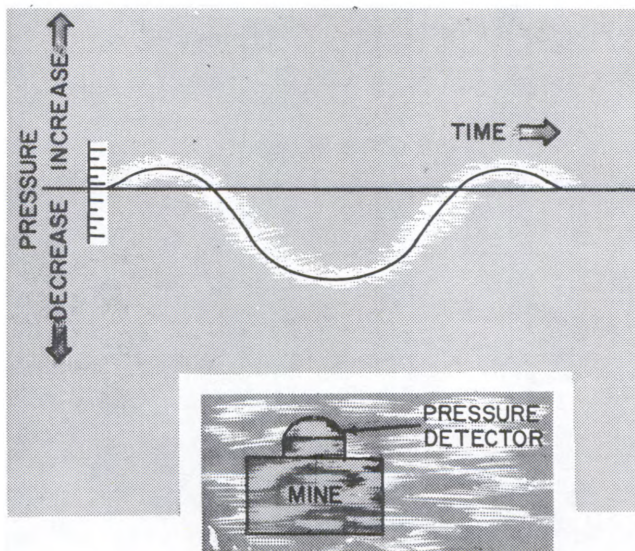
Aerodynamic pressure-sensing devices are used to measure altitude, air speed, air density, barometric pressure, and ram or impact pressure.

In guided missile control, where the missile response changes with the impact effect of the air molecules, a direct measurement of air molecule density rather than pressure is required. Methods of measurement of aerodynamic pressure include: sealed flexible bellows whose dimension will change relative to ambient pressure, as in an altimeter or barometer; a bellows connected to an open tube to measure ram pressure, as in an air speed indicator; an electric altimeter cell, as one in which the static molecular density determines the cooling rate and thus the temperature and resistance of a heated platinum wire.

Hydrostatic pressure-sensing devices are used to measure ocean depth and ship speed, and to detect ship or submarine pressure signatures. In addition to devices similar to those used for aerodynamic pressure-sensing, hydrostatic pressure is also measured by piezoelectric devices similar to those used for hydrophones.

As a ship or submarine travels through the water, the displacement and flow of water around the hull causes variations in the water pressure which are detectable even at a considerable depth. The pressure variation as a ship passes a given point is shown schematically. This curve is referred to as the pressure signature of the ship. Pressure-influence firing mechanisms of mines are designed to detect such pressure signatures and to differentiate them from other variations in water pressure.

Pressure-sensing devices for depth determination are also used by hunting torpedoes which are designed to circle progressively at various depths until auxiliary sensing devices detect the presence of a target.



pressure signature of a ship

magnetic sensors

Magnetic sensors are used primarily for the detection of submarines and ships. In fact, the only widely used nonacoustic device for submarine detection is a magnetic sensing device called the magnetic-anomaly detector (MAD). The sensing element of MAD is a magnetometer, which is essentially a multiturn coil on a saturable core. The coil produces an induced emf when exposed to varying magnetic fields.

Because of the high magnetic permeability of a steel-hulled ship or submarine, the earth's magnetic flux lines tend to concentrate through the ship, disturbing the normal earth's magnetic field pattern in that area. In addition, every steel ship is magnetized to some extent (in spite of possible reduction with degaussing coils), so that each ship is equivalent to a magnetic movement dipole, which will also distort the earth's field. Such disturbances or anomalies in the earth's magnetic field can be detected by a magnetometer passing through this varying field up to a range, at present, of 1000 to 1500 feet.

Most MAD installations are carried in aircraft or helicopters to obtain maximum rate of search. Dipped and towed MAD units are being considered to increase the range of detection.

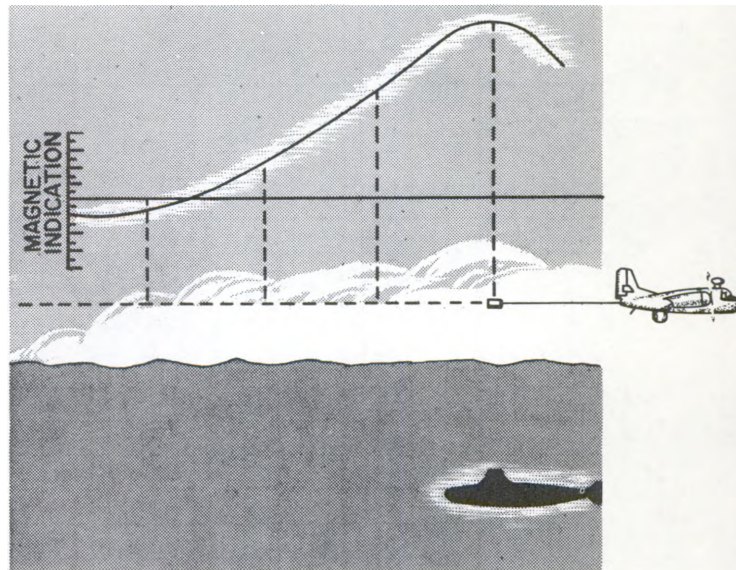
Magnetic-influence firing mechanisms are also used on mines. A search coil is used to detect the magnetic signature of a passing ship.

Magnetic detection devices on the ocean bottom are also used in harbor defenses in the form of large induction loops and coils.

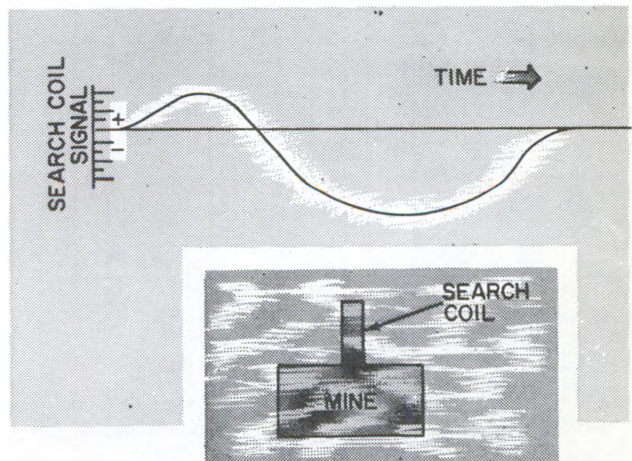
Many new and different sensing methods are as yet untried or have not yet been conceived. Among those in use or under development, and not previously listed in this section, are:

Thermal sensors (for temperature measurements); radiation sensors (for detection of nuclear waste from nuclear subs); cosmic ray and ionization sensors (for possible detection purposes); electrostatic sensors (used for close range detection by missiles, torpedoes, and subs); and television (used by drones for battlefield surveillance and for possible underwater detection).

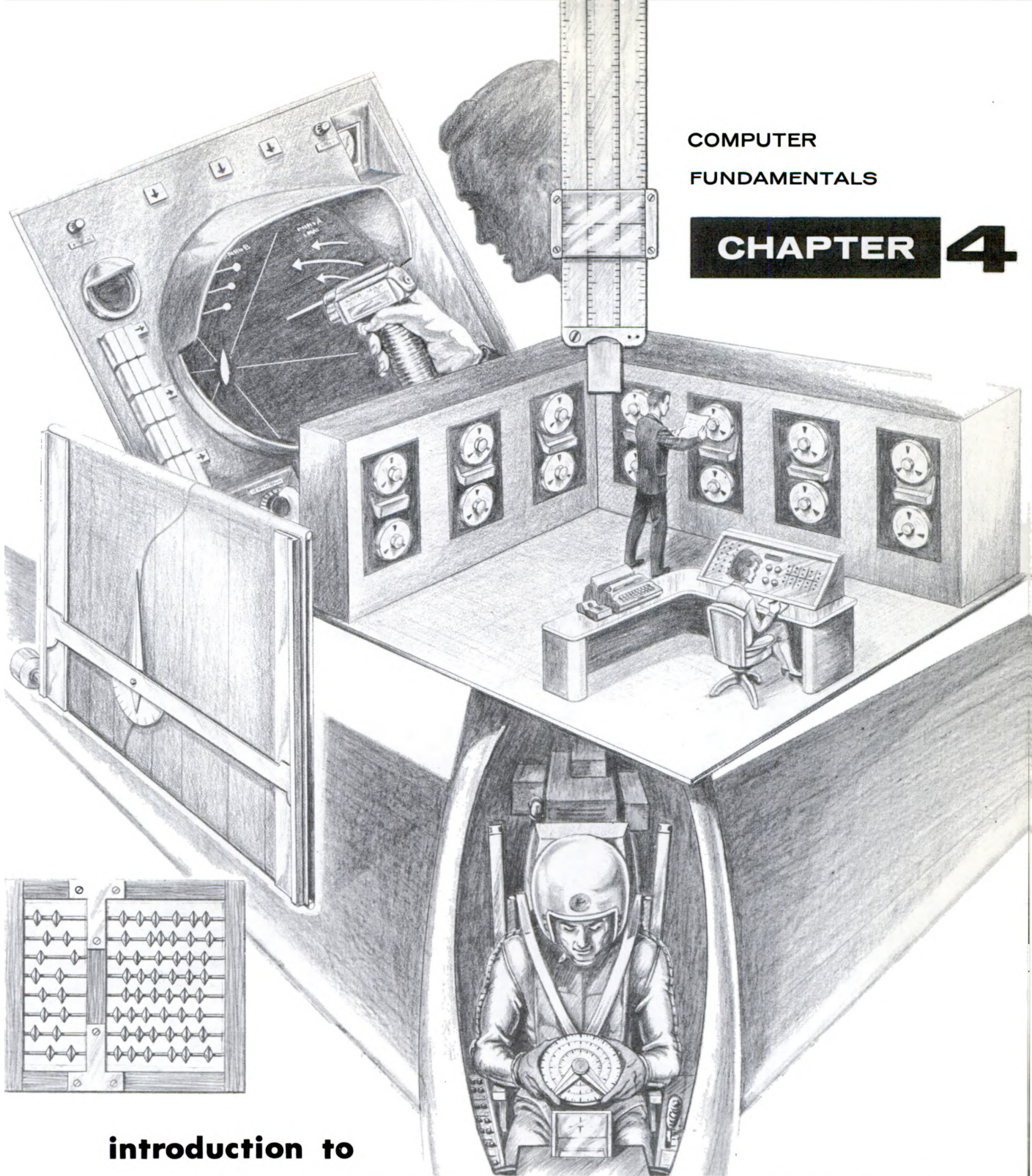
The possibilities and applications for sensors obviously are not limited.



magnetic anomaly detector



magnetic signature of a ship



introduction to

COMPUTER FUNDAMENTALS

The preceding chapters have explained the need for data processing in the operation, design and support of weapons systems. Tactical fire control requires storage of a great deal of information concerning the location and movement of friendly and enemy units, and an intelligible display of this data; control of information gathering equipment, etc.; and selection and control of weapons. The design of weapons systems entails complex mathematical calculations, simulation, inventory and production control, testing and communications. Logistic support and maintenance involves keeping track of personnel, equipment, and parts. The large volume of information to be processed, and the brief time available to handle it, make the use of high-speed data processing equipment essential.



The many variables which affect the chance of scoring a hit, and the necessity of rapid and accurate calculations in weapons control systems were discussed in Volume II. Other weapons system tasks as well can be performed or aided by computers. The search, detection, and classification phases of the fire control problem are becoming subject to a higher degree of automation. The quantity and complexity of weapons systems components (particularly electronic equipment) is increasing the importance of devising automatic checkout and testing. This section outlines the capabilities of computers, some methods and devices used, and the characteristics of various types of computers.

Data processing equipment is a group of devices, each capable of performing a mathematical operation on data furnished to it, and of producing the results in a usable form. The term "mathematical operation" is not limited to such obvious processes as multiplication, differentiation, etc. Any process which can be described in

a mathematical notation can be mechanized in a computer. This includes analytic, arithmetic and logic operations, but excludes such processes as judgment, induction, conjecture and generalization.

The term "computer" is generally applied to any data processing device capable of distinguishing between different datum, whether its function is principally mathematical or not. Devices which translate data from one form to another, or store data and produce it on demand without regard to the nature of the data, are called data processing equipment.

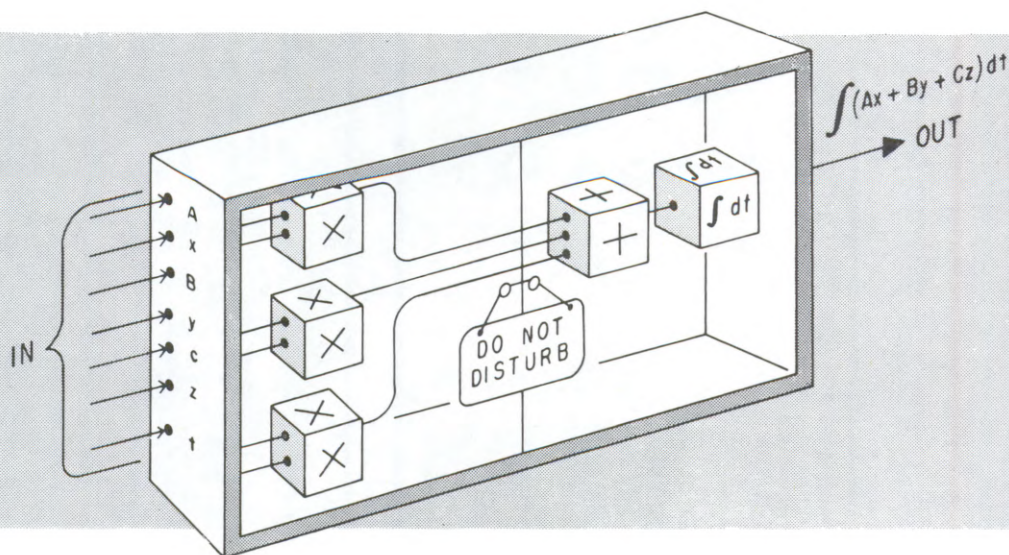
Induction refers to the process of reasoning from a particular case to the general case; i.e., recognizing that a particular law is valid for a finite number of cases and inferring that it is valid for all cases. The process called "mathematical induction" is misnamed, and is not induction at all.

SPECIAL AND GENERAL

The various devices which make up a computer must be connected in the proper sequence to produce the desired result. They may be connected permanently

a special purpose computer.

i.e., one intended to solve one problem, or a predetermined small number of problems, is usually connected permanently.



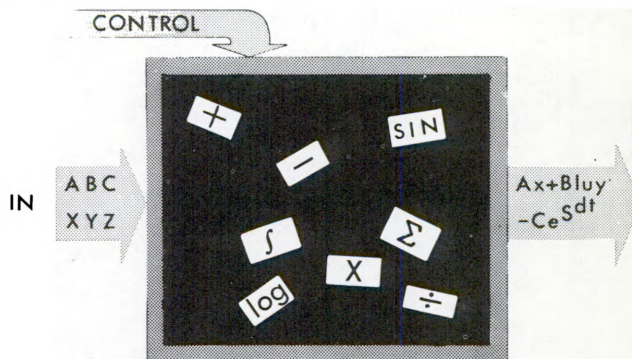
This is a generalization which is seldom completely true in practice, however. Often there is some provision for altering connections in special purpose computers. For example, a computer which guides interceptor aircraft on a pass at a target allows the pilot to vary the computer set up to select one of two basic types of approach courses. A ballistic missile guidance computer will solve only the equation for a ballistic trajectory, but can be made to guide the missile to any target, within range. Both are examples of special purpose computers, because they can only be used for one basic problem, yet some

variations are permitted in the arrangement of components. In a general purpose computer, a group of components used jointly to perform some frequently occurring operation, may be permanently connected, and used as a unit in handling problems requiring this particular operation. The reader is reminded that the problems which arise in weapons systems are confined to a particular field. Therefore, while these problems could be handled by a general purpose computer, any computer designed specifically for weapons systems will usually be a special purpose computer. This is because a gen-

In order to retain some order in this discussion, computers are classified according to purpose (general or special), and according to whether they use analog or digital principles (defined later). This does not mean that all computers can be assigned to a category with no "gray areas" in between; nor that all classes are of equal importance in weapons systems. Computers used in weapons systems are designed specifically to solve problems arising in weapons systems, and therefore are usually special purpose. Paradoxically, a large percentage of the chapter is more applicable to general purpose computers. This is primarily because general purpose computers of a particular type (analog or digital) are very much alike, but special purpose computers occur in a wide variety of forms which may have little in common with one another.

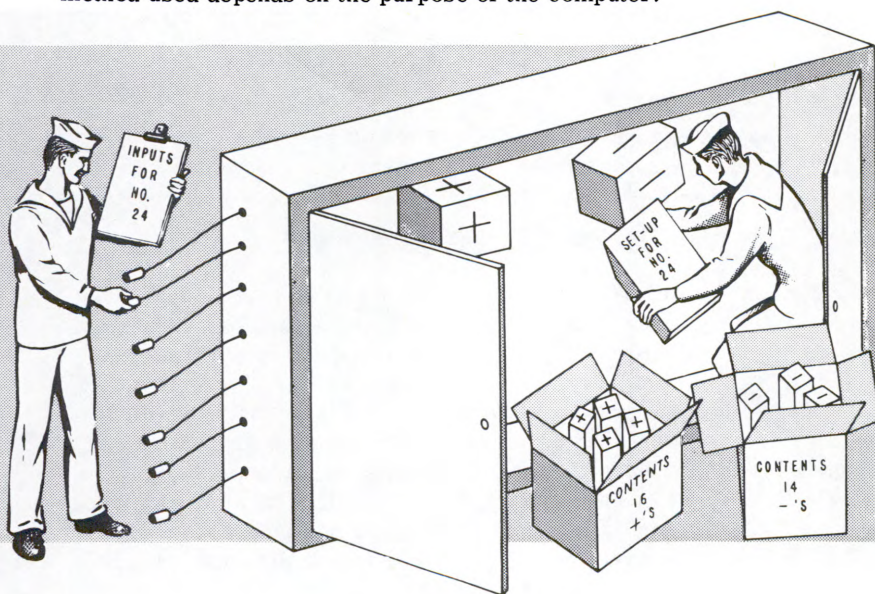
To the extent that a computer can be discussed intelligibly without resorting to classification, it may be considered as a black box, with input data, input controls (which may be as simple as an on-off switch), and output

data. The box consists of a group of devices connected in a manner which will give the desired output. The nature of the device determines whether it is classified as analog or digital, and the manner in which the connections are established determines whether it is general or special purpose.



PURPOSE COMPUTERS

or left unconnected to be set up for each problem. The method used depends on the purpose of the computer.



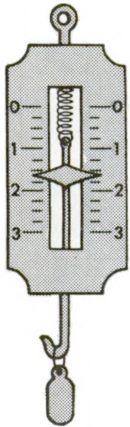
a general purpose computer,

i.e., one intended for scientific or general business use, is normally set up for each problem.

eral purpose computer capable of solving a given problem will cost more than a special purpose computer designed to solve the same problem. An exception might be found in a computer designed to handle a problem of such great complexity that the computer must include the ability to handle a wide range of lesser problems in order to be capable of handling the specific problem for which it is intended. Here it is necessary to distinguish between problems of a recurring operational nature, and those of a more singular or strategic nature. In the first class, weapon

control, navigation, etc., the problems are not this complex. In the second class, which includes such problems as simulation of battles to evaluate weapons, or to form tactical doctrine, the problems do not usually arise with sufficient frequency to warrant designing a computer specifically for them. Existing general purpose computers are used instead. While general purpose computers are used to solve problems directly related to weapons systems, most computers which are encountered as part of a weapons system are special purpose.

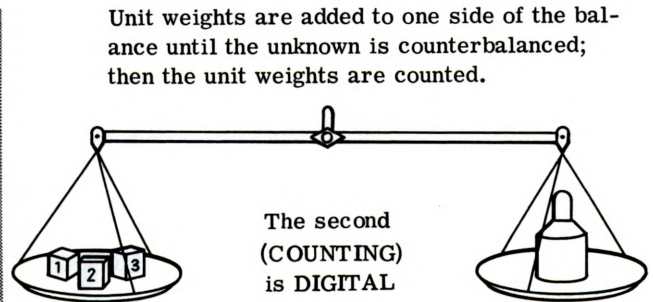
ANALOG AND DIGITAL COMPUTERS



One of two methods is normally used to determine the magnitude of a variable: measurement or counting. For example, either a spring scale or a balance could be used to determine the weight of an object.

The deflection of the spring scale is proportional to the weight, and therefore measures weight directly.

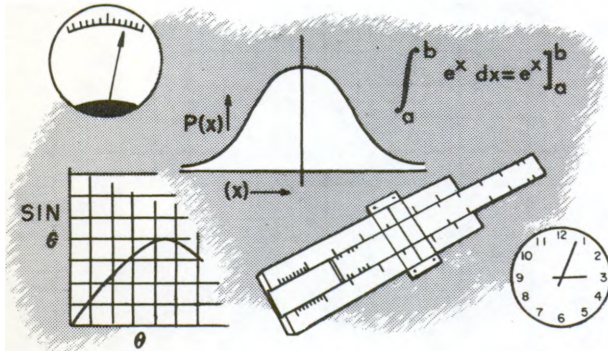
The first
(MEASUREMENT)
is ANALOG



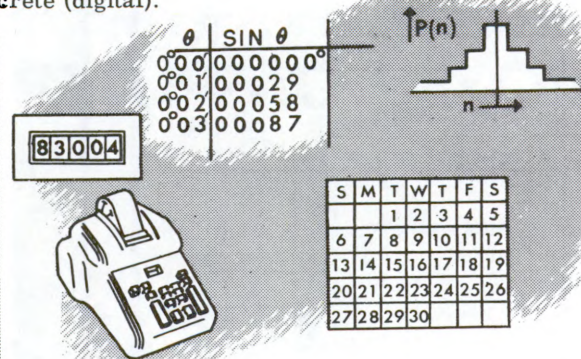
Unit weights are added to one side of the balance until the unknown is counterbalanced; then the unit weights are counted.

The second
(COUNTING)
is DIGITAL

The difference is fundamental, corresponding to the mathematical distinction between the continuous (analog) and the discrete (digital).



The analog device recognizes a variable as a whole quantity: displacement of a pointer, rotation of a shaft, voltage in a circuit, etc.



The digital device recognizes a quantity as a number of basic units: number of weights on a balance, days on a calendar, etc.

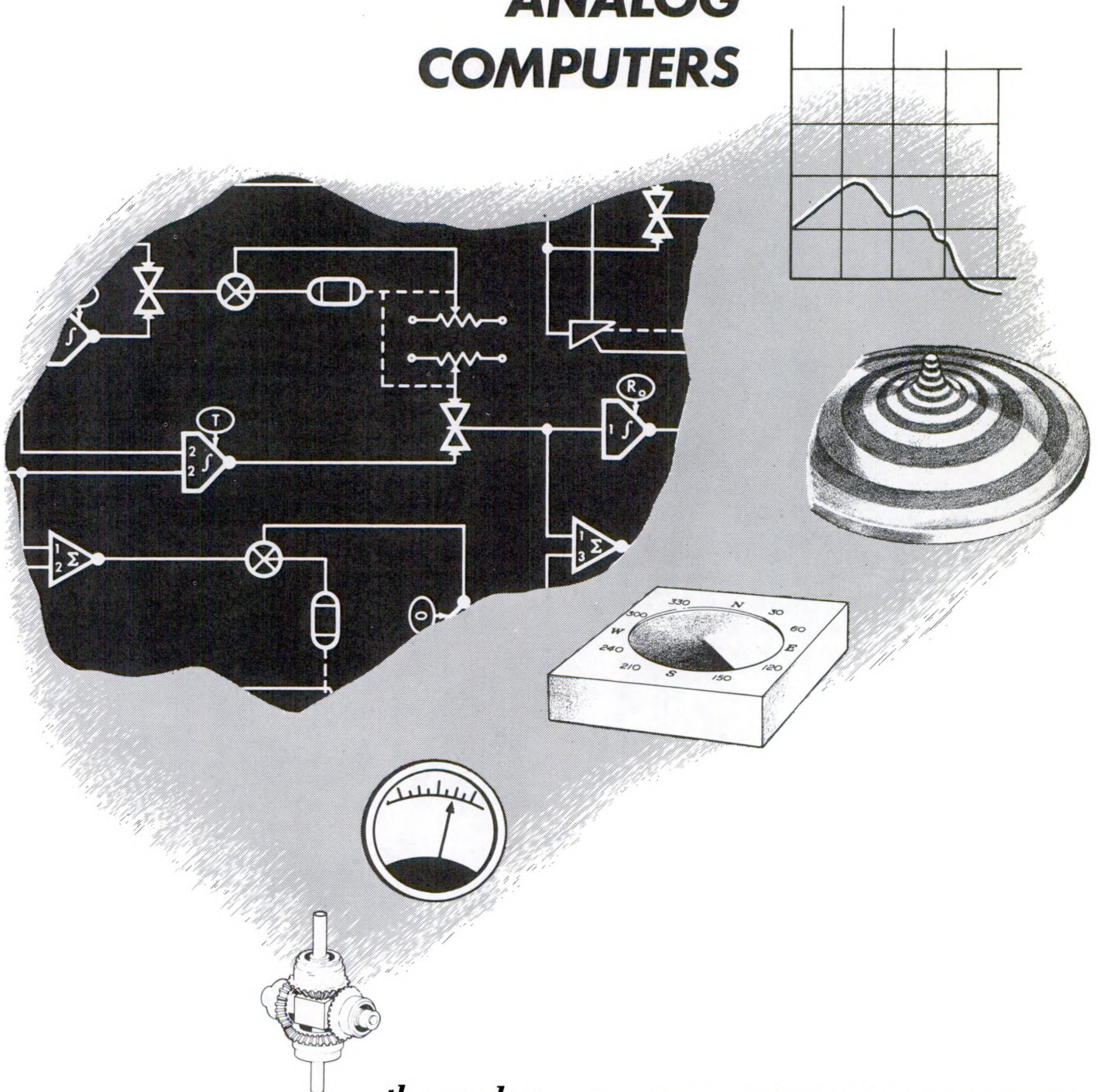
Note that the analog device will continuously measure a changing variable. If the object to be weighed were a leaking bucket of water, for example, the spring scale would show the weight of the bucket and water at all times. The balance, however, would give a value which is correct only for an instant. It is also important to realize that the analog device will read any fraction of a unit. The limit of precision is determined by how precisely the scale can be read. The limit of precision of the digital device depends only on the size of the smallest basic unit used (smallest weight in this case).

Just as the mathematical distinctions between the continuous and the discrete leads to such different techniques as integration and summation, or graphs and tables, analog and digital computing methods involve major differences in techniques and capabilities.

Each technique offers some advantages and incurs some disadvantages. It would seem reasonable to construct a computer by combining analog and digital devices, each operation being performed by the type of device best suited to it. However, using both techniques together requires conversion of data from one form to the other and

back. While this can be done with little difficulty, the necessary added equipment often negates the advantage gained. Furthermore, combining two techniques often results in losing some advantages of each while compounding the disadvantages. For example, digital devices are often capable of greater precision than analog devices while analog devices are frequently more tolerant of adverse temperature and humidity. If both types are combined in the same machine, the accuracy of the result cannot be greater than that of the least accurate device, while sufficient environmental control must still be maintained for the least tolerant device. Sufficient advantages remain so that computers combining both techniques are frequently used. They remain a combination of two distinctly different techniques, however. Since analog and digital techniques differ widely in the fundamental manner in which they operate, their devices, characteristics and requirements have very little in common. Therefore, the discussion in this text has been separated into two parts, concluding with a brief discussion of combined analog and digital techniques.

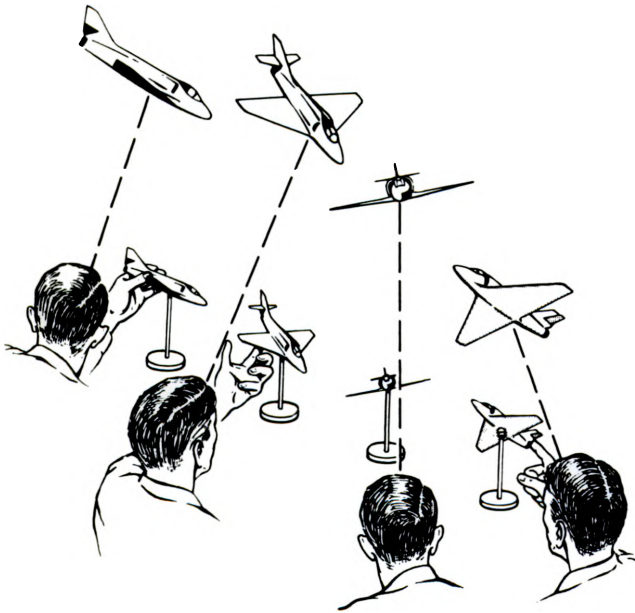
ANALOG COMPUTERS



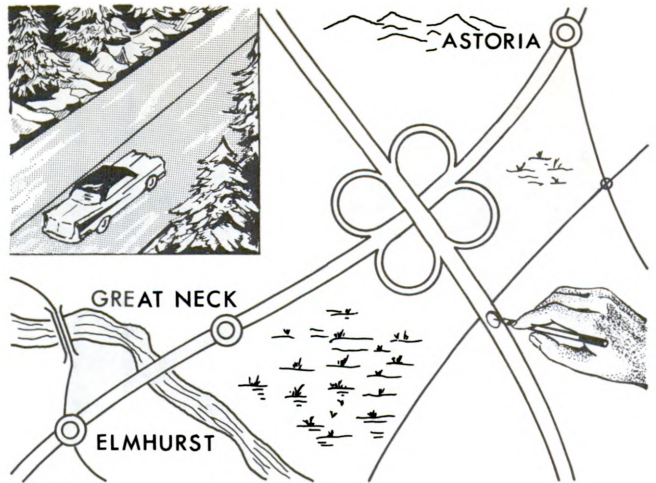
the analog

An analog is any device which duplicates one or more features of a subject. The analogy, i.e., area of duplication, may be physical (e.g., a scale model of the subject), or purely mathematical. The purpose of any useful analog is to allow the person using the analog to observe or measure some feature which is not conveniently available on the subject.

The control of weapons may require the knowledge of various angular parameters of a target, but a target is not normally accessible for direct measurement.



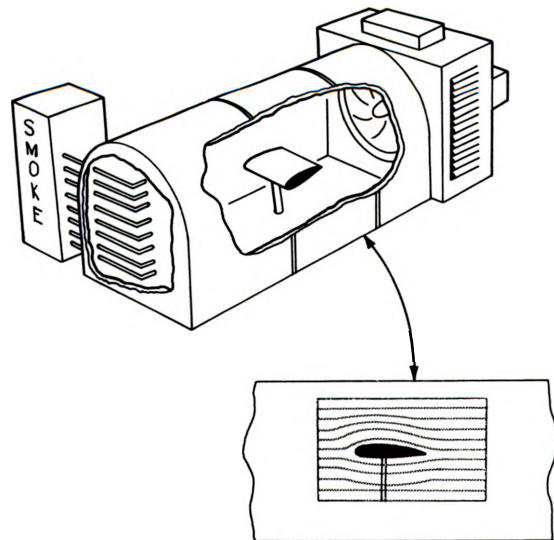
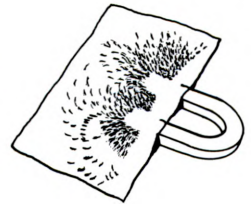
By using an analog of the target, a model which is made to assume the same attitude as the target, the heading, rate of turn, etc., can be measured by instruments attached to the model. While extremely simple, the characteristics of all analog computing devices are present. The analog duplicates some important features of the subject, and is created to make measurements relating to the inaccessible or inconvenient subject.



A road map is another "scale model" analog. Measurements of distance can be easily made by means of any device which will measure distance on the map. The map is an analog of the terrain, and the mileage measuring device is an analog of a vehicle moving from one point to another.

An interesting application of analogs using the "scale model" idea is in the area of depicting fictitious or hypothetical concepts such as contour lines, field lines, equipotentials, streamlines, etc.

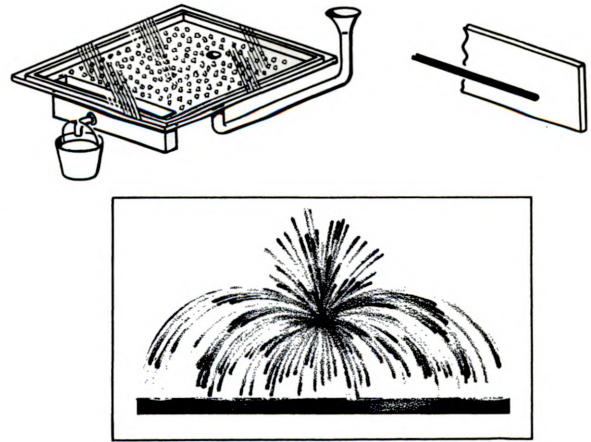
Since these exist only in the imagination, some analogous model is indispensable if measurements are to be made. An example is the use of iron filings to show the field lines around a magnet.



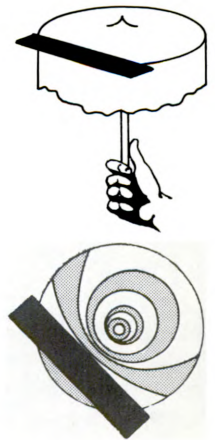
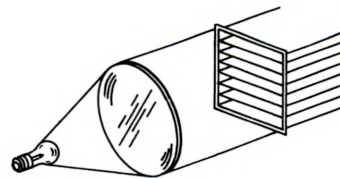
One of the simplest of these analog devices is used to examine streamlines in air flowing past an obstacle. The device is much like a wind tunnel, except part of the air entering the tunnel is drawn through a line of small holes from a smoke-filled chamber. With no obstruction in the tunnel, the smoke would form a series of parallel lines running the length of the tunnel. With an obstacle in the tunnel, the cross section of a wing, for instance, the smoke will follow the "streamlines", making them visible to an observer.

Another example of a physical analog is the model aircraft or ships used in wind tunnel or model boat basin tests. By duplicating the geometry of the actual subject, the analog will also duplicate the aerodynamic or hydrodynamic characteristics of the subject. By reducing the size, measurement is simplified.

A similar device is the fluid mapper which is used to study the streamlines of a flowing liquid in a plane. This device consists of a flat bed covered with a sheet of glass so the liquid is constrained to flow in the plane between the bed and the glass. A water soluble dye, (e.g., KMnO_4) in crystalline form, is scattered around the bed, and sources and sinks are provided for water to enter and leave. As the water flows, the dye is dissolved, and leaves visible traces showing the streamlines. Since the field lines of an electric field are analogous to the streamlines in the plane flow of a liquid, the fluid mapper can also be used to depict electric fields. The field lines of a charged line near a conducting plate are shown as they appear on a fluid mapper.

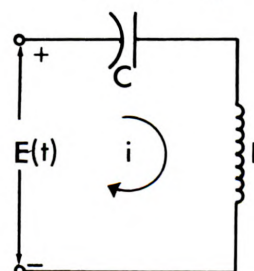
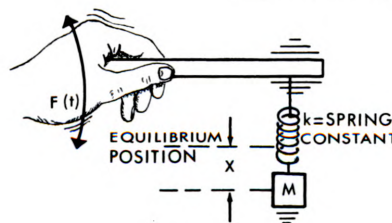
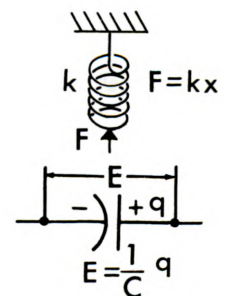
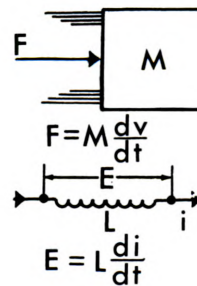


The deflection of an elastic sheet under tension is analogous to the potential in an electric field. Therefore, a rubber sheet can also be used to study electric fields. The height above the reference plane is analogous to the potential, and the lines of equal height (contour lines) are analogous to the equipotential lines. The sheet is stretched over a frame. Charges are simulated by pushing up the sheet with a rod, to a height analogous to the quantity of charge. A conducting plate can be simulated by laying a heavy plate on the sheet. (The important electrical property of a conducting plate is that it forces the potential to be zero at the surface of the plate, just as the metal plate forces the height of the sheet to be zero at the edge.) The contour lines are created by projecting a light beam through a slide containing closely ruled lines. This creates a series of equally spaced, alternately light and dark horizontal planes of light, which are projected parallel to the plane of the sheet. The equipotentials of a charged line near a conducting plate are illustrated. Note that the equipotential lines are perpendicular to the field lines shown in the illustration of the fluid mapper.



All these analogs are based on a physical similarity between the model and the subject. Analogs based on a mathematical similarity can normally be used in more widespread applications.

For example, the equation for the force on a mass is $F = M \frac{dv}{dt}$, and the equation for the voltage across an inductor is $E = L \frac{di}{dt}$. Therefore, an inductor may be considered analogous to a mass, with inductance equivalent to mass, voltage equivalent to force, and current equivalent to velocity. Similarly, the equation for the force due to the displacement of a spring is $F = kx$, where k is the spring constant and x is the displacement. The equation for the voltage across a capacitor is $E = q/C$, where C is the capacitance and q is the charge. Therefore, a capacitor is analogous to a spring, with voltage equivalent to force, charge equivalent to displacement and capacitance equivalent to the reciprocal of the spring constant. Note that the two situations are compatible; in both cases, voltage is equivalent to force. Furthermore, velocity, which is equal to dx/dt , is equivalent to current, which is equal to dq/dt , and charge (q) is equivalent to displacement (x). Therefore, the two devices, an inductor and a capacitor, can be connected to simulate a weight on a spring. The dynamic behavior of the mechanical system under various conditions can be measured by creating analogous conditions in the electrical system.



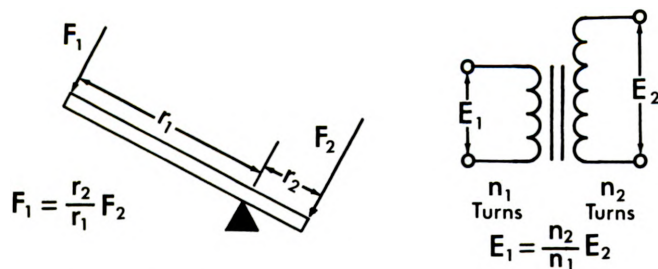
$$M\ddot{x} = -kx + F(t)$$

$$M\dot{v} = -k \int_0^t v dt + F(t)$$

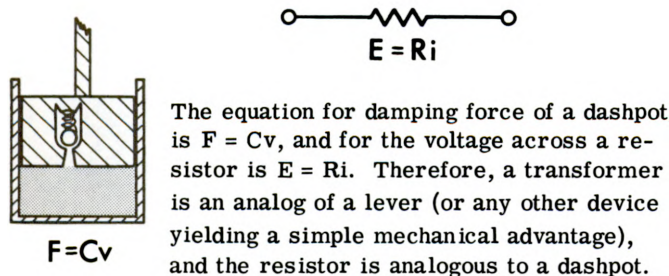
$$L\ddot{q} = -\frac{1}{C}q + E(t)$$

$$Li = -\frac{1}{C} \int_0^t i dt + E(t)$$

Many more analogous mechanical and electrical devices are available, retaining the same correspondence between mechanical and electrical quantities.

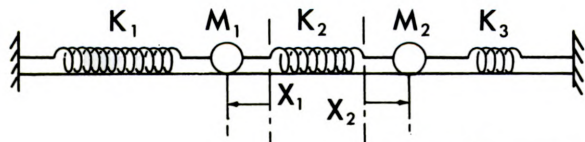


The relation between forces on a lever is $F_1 = \frac{r_2}{r_1} F_2$, and the relation between ac voltages across a transformer is $E_1 = \frac{n_2}{n_1} E_2$.



It is not necessary that force always be represented by voltage, velocity by current, etc., but if several devices are to be combined in the same circuit, it is essential that they be consistent with one another, as is the case with all the examples given here.

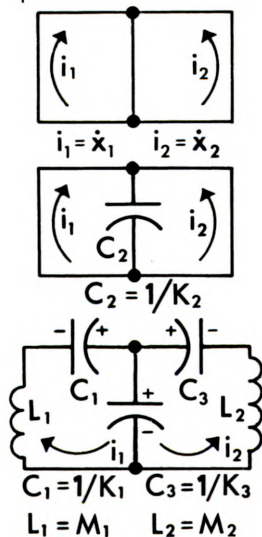
An electrical analog of two masses connected by three springs, for example, may be constructed directly without knowledge of the differential equations describing the behavior of the spring mass system.



Since there are two displacements involved, and therefore two velocities, two current loops are required.

The center spring is affected by both masses; therefore the capacitor which represents it must go in the center arm of the circuit.

Adding an inductor in each loop for the masses, and a capacitor for each remaining spring completes the analog.



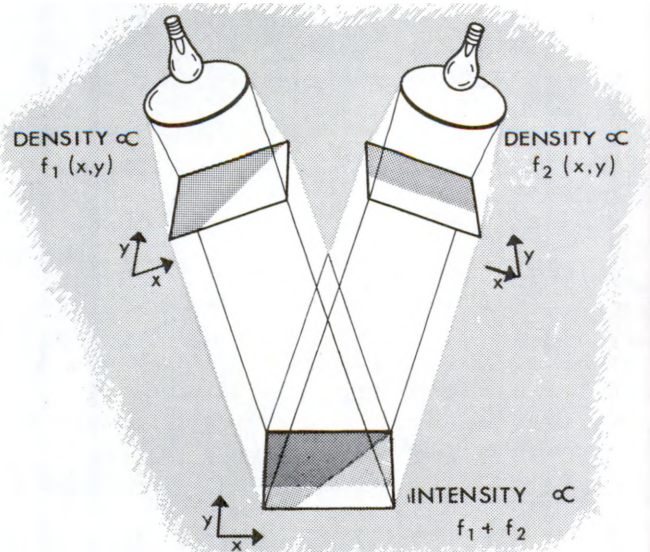
To check the sense of the quantities, note that positive velocities of the masses tend to add in their effect on the center spring. Since current is equivalent to velocity, positive currents i_1 and i_2 should add in the center arm, as shown. Positive velocities tend to expand the center spring, so a charge on C_2 as shown in the diagram represents an expanded spring. Positive velocities tend to compress the remaining springs, so a charge on capacitors C_1 and C_3 opposed to positive current flow represents expanded springs.

As a practical use of the analog, consider the problem of studying the effect of displacing both springs slightly from equilibrium and then releasing them. This situation can be placing an initial charge on the capacitors equal to the initial displacement of the masses.

An ammeter in each loop will show the velocity of each mass, and since the charge on a capacitor is equal to CE , the voltage across the capacitors, divided by C , will show the displacement of the masses.

Analogues are not restricted to the relations given here. It is also possible to form analogues where, for example, force would be analogous to current, and displacement to voltage, or many other combinations. The only requirement is that the analogies be consistent so that the devices are compatible and can be interconnected.

Analogues are not restricted to electrical or mechanical devices. They have been constructed based on laws of chemistry, hydraulics, acoustics and optics.



For example, optical devices are frequently used to solve problems involving the sum or product of several functions of two variables (e.g., $F(x,y) = f_1(x,y) + f_2(x,y) + f_3(x,y)$, etc.). A transparency is made for each function, with the density at each point equal to the value of the function at that point. By exposing a single sheet of photographic paper to light projected through each transparency in turn, an image whose brightness is equal to the value of $F(x,y)$ at each point is obtained.

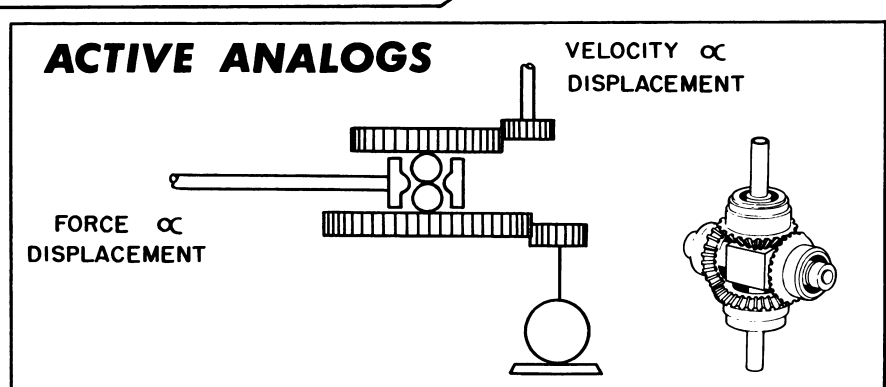
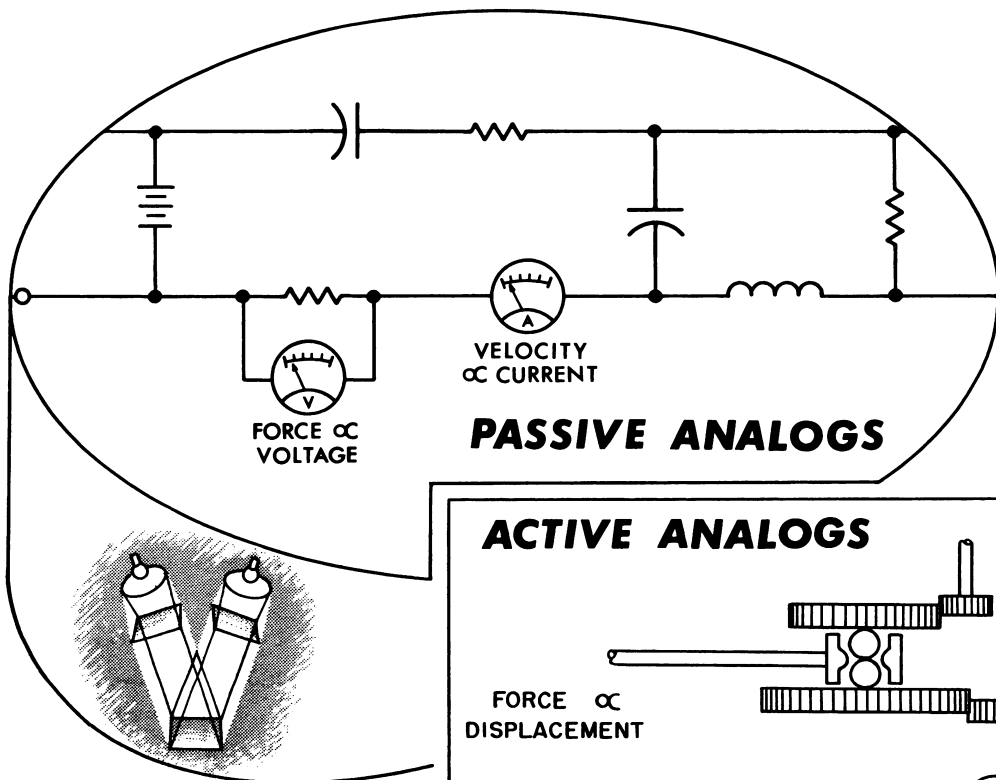
ACTIVE vs PASSIVE ANALOGS

All the analogs discussed so far have depended upon a one-to-one substitution of variables. Height was substituted for potential, voltage for force, current for velocity, etc. Each variable of the analog system represented one (and only one) variable of the subject system. Furthermore, the analog duplicated the subject system as a whole. For complex systems of many variables, particularly those involving more than one degree of freedom, this type of analog is difficult to design.

By using a type of analog made up of a number of basic elements, each capable of performing a mathematical operation, i.e., integration, multiplication, addition, etc., analog computer systems can be designed with little difficulty. Using this type of construction, the analog can be built up a piece at a time, and it is not necessary to use a different analog variable for each subject vari-

able. As a rule, it is most convenient to use one variable, or a small number of variables, throughout the analog system, to represent all variables of the subject system. Thus, voltage in the analog system may represent velocity at one point, and displacement at another point. Those analog computers which represent each subject variable by a different analog variable are called "passive" analogs, while those which use one analog variable to represent several subject variables are called "active" analogs. (The terms "passive" and "active" stem from the fact that the latter type usually contains power amplifying (active) devices, while the former type is made up of devices which do not amplify power (passive devices). This is a secondary characteristic, however, and is only incidental to the fundamental difference.)

Where the passive analog would represent velocity as a current (dq/dt), and displacement as charge (q), the active analog would use the same variable to represent both velocity and displacement, in the first case as the input to an integrator, and in the second case as the output of the integrator.

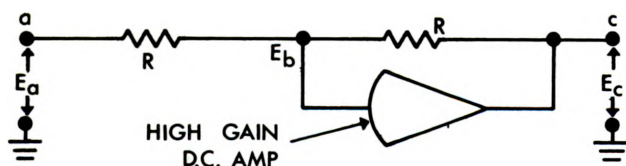


Several mechanical analog devices capable of addition, multiplication, formation of sines and cosines, etc., were described in Volume I. Some electronic and electromechanical devices are described here.

ANALOG DEVICES

OPERATIONAL AMPLIFIERS

The operational amplifier consists of two resistors and a very high gain d.c. amplifier.



The important property of the high gain amplifier is that a very small positive voltage at point b is amplified to a high negative voltage at point c. The high gain amplifier also presents a very high resistance (typically 10^5 to 10^6 ohms). Therefore, for practical purposes, it may be assumed that no current flows through the amplifier, and the current through each resistor will be equal. From Ohm's law, the voltage between points a and b, is equal to IR .

$$E_a - E_b = IR \quad (1)$$

The voltage between points b and c is also equal to IR .

$$E_b - E_c = IR \quad (2)$$

subtracting (1) from (2):

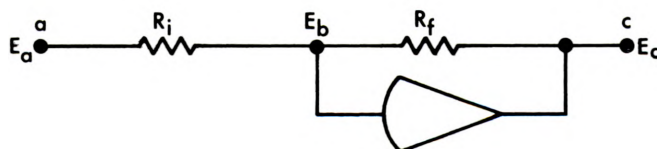
$$E_b - E_c - E_a + E_b = IR - IR$$

$$2E_b - E_c - E_a = 0 \quad (3)$$

$$E_b = \frac{E_a + E_c}{2} \quad (4)$$

In other words, the voltage at point b is the average of E_a and E_c . If a positive voltage is applied at point a, it will tend to increase the voltage at point b. This voltage will be amplified by the high gain amplifier to create a negative voltage at point c. When E_c is equal to $-E_a$, the voltage at b will again be zero. (It will not be exactly zero, since there must be some voltage present at point b to be amplified. The gain (A) of the amplifier is so great, however, that about 1/1,000,000 volt is required at point b to produce -1 volt at point c.) The function of the high gain amplifier, therefore, is to produce a sufficient voltage at point c so that the voltage at point b remains very close to zero. If both resistors are equal, this will make $E_c = -E_a$. The first resistor is called the input resistor (R_i), and the second is called the feedback resistor (R_f).

If $R_f \neq R_i$, the voltage from a to b, (E_a since $E_b = 0$) will be equal to IR_i , and the voltage from b to c ($-E_c$) will be equal to IR_f .

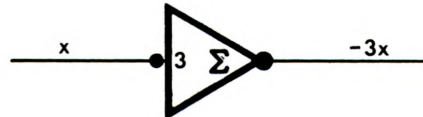


Therefore:

$$\frac{-E_c}{E_a} = \frac{IR_f}{IR_i} \quad (5)$$

$$E_c = \frac{-R_f}{R_i} E_a \quad (6)$$

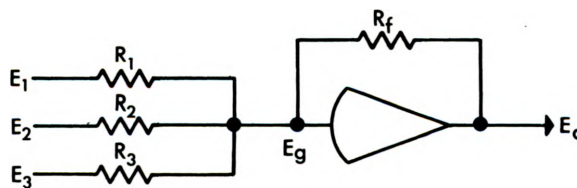
The operational amplifier multiplies the input voltage by a constant R_f/R_i , and inverts it. When the operational amplifier is represented symbolically, the factor R_f/R_i is usually included in the symbol. The symbol illustrated here represents an operational amplifier with $R_f/R_i = 3$. The small dot at the apex of the symbol indicates that the sign of the function is changed.



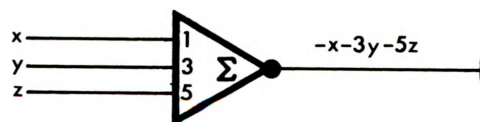
addition

Substituting several input resistors for R_i , equation (6) becomes:

$$E_o = -\frac{R_f}{R_1} E_1 - \frac{R_f}{R_2} E_2 - \frac{R_f}{R_3} E_3 \quad (7)$$



In this form, the operational amplifier functions as an adder, while simultaneously multiplying each variable by a constant. Again, the resistance ratios are specified on the symbol.



integration

An operational amplifier can be made to integrate by substituting a feedback capacitor for the feedback resistor. Equation (1) then becomes:

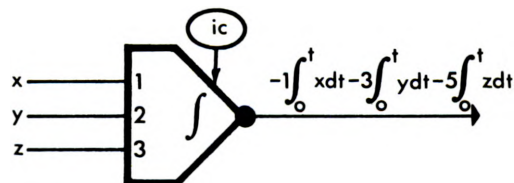
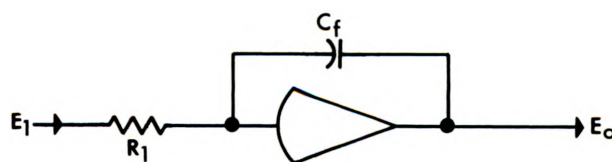
$$\frac{E_g - E_1}{R_1} = \frac{dq_f}{dt} = C_f \frac{d(E_o - E_g)}{dt} \quad (8)$$

since $E_g = E_o/A = 0$:

$$-\frac{E_1}{R_1} = C_f \frac{dE_o}{dt} \quad (9)$$

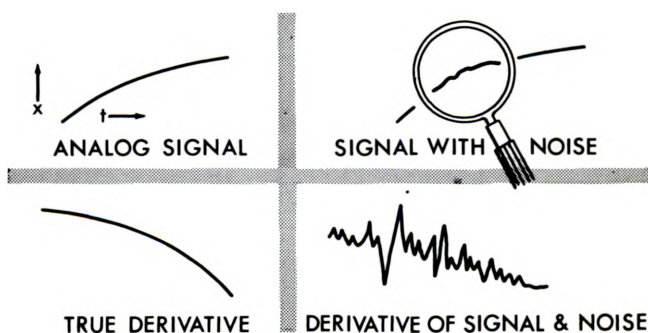
$$E_o = -\frac{1}{R_1 C_f} \int_0^t E_1 dt + E_{ic} \quad (10)$$

where E_{ic} is the initial voltage across C_f at $t = 0$. Several input resistors can be used in this configuration also, allowing the operational amplifier to add several integrals.



differentiation

The operational amplifier with an input capacitor and feedback resistor will differentiate, but electronic differentiators are seldom practical because of the presence of noise. The actual input to an electrical device is the sum of the signal plus the noise. The amplitude of noise can be kept small compared to the amplitude of the signal, and since it is random (negative as often as positive) the contribution to the integral is negligible. The rate of change of the noise is usually much greater than that of the signal, and its contribution to the derivative is also greater than that of the signal.

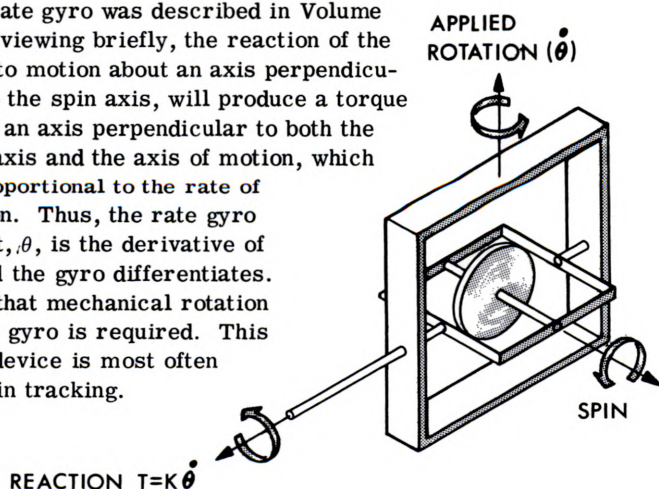


Fortunately, there is no real need for differentiators, since any expression can be differentiated analytically with little difficulty, resulting in a differential equation which can be solved by integrators. Electromechanical differentiators are often practical (even though electronic differentiators are not), if the inertia of the mechanical system is large. This is due to the fact that the inertia

of the mechanical components makes them less susceptible to noise (erratic motion), and the mechanical differentiator usually has a frequency response too small to respond to the effects of the noise. Two such differentiators used frequently enough in weapons systems to warrant discussion are the rate gyro and the tachometer.

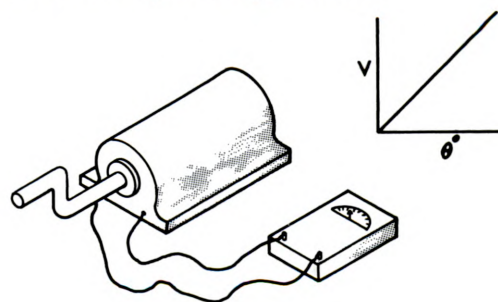
rate gyros

The rate gyro was described in Volume I. Reviewing briefly, the reaction of the gyro to motion about an axis perpendicular to the spin axis, will produce a torque about an axis perpendicular to both the spin axis and the axis of motion, which is proportional to the rate of motion. Thus, the rate gyro output, $\dot{\theta}$, is the derivative of θ , and the gyro differentiates. Note that mechanical rotation of the gyro is required. This type device is most often used in tracking.



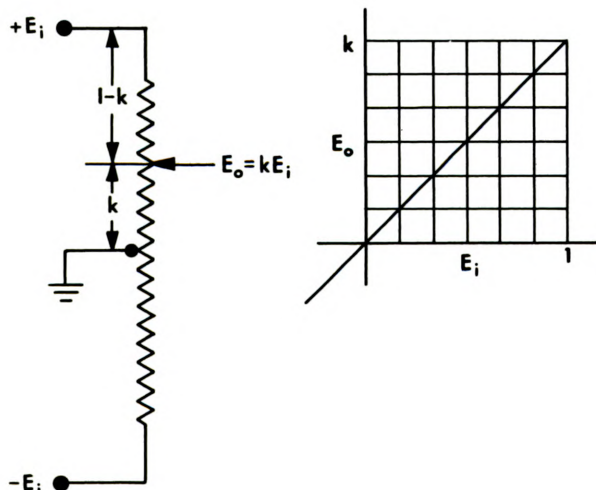
tachometers

The most common form of tachometer is a generator whose output voltage is proportional to the rate of rotation of the rotor. Note that only the input shaft must be rotated, rather than the entire device, making the tachometer more useful in general applications.

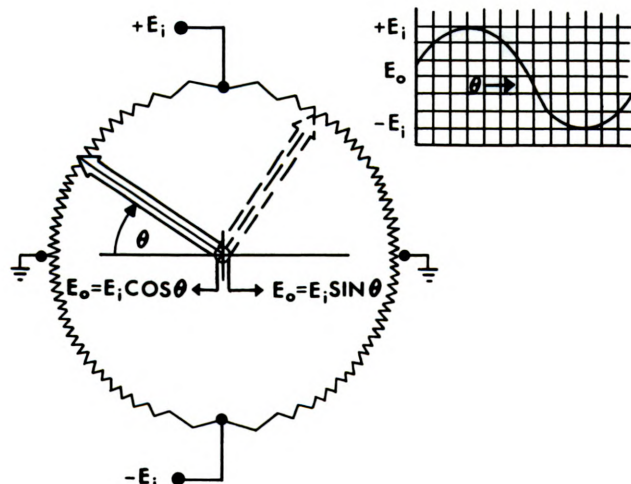


POTENTIOMETERS

ter is the most convenient device for generating a voltage by a constant. A linear potentiometer will multiply a voltage by any constant with a magnitude less than one. A nonlinear potentiometer, i.e., one in which the resistance is a nonlinear function



of the displacement, can be used to generate any desired function of a variable, e.g., $\sin x$, $\cos x$, $\log x$, e^x , etc., when the independent variable (x) is represented by a linear or rotational displacement.



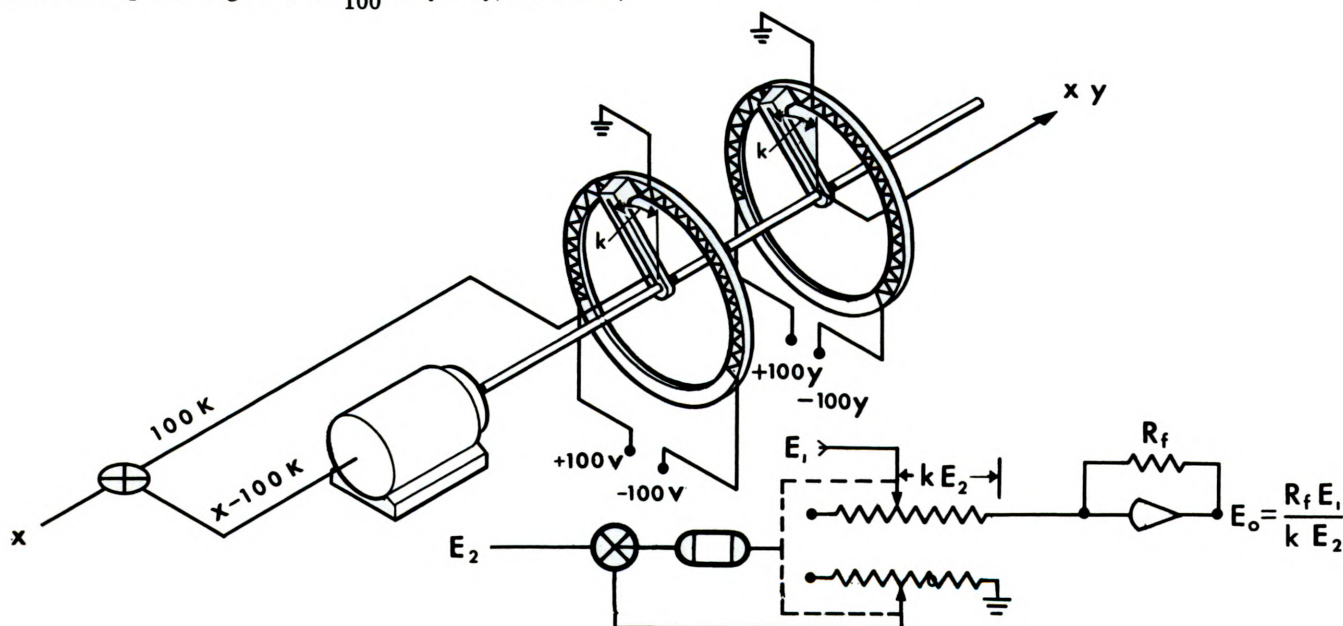
SERVOMECHANISMS

The servomechanism is generally used to represent a variable by a linear or rotational displacement. As described in Volume I, a differential and servo motor can be used to force a mechanical displacement determined by the input signal to the differential. Arranged as shown here, the servomechanism will move the potentiometer wipers until the response (x) is equal to the input (x). This will occur when the displacement of the wipers is equal to $x/100$. Since the two potentiometer wipers are driven to the same position, the displacement (k) of the second wiper will also equal $x/100$, and the output voltage will be $\frac{x}{100} 100y = xy$; therefore,

the device is a multiplier.

The ability of a servomechanism to position a potentiometer wiper to a point proportional to a voltage can also be used to divide. The output voltage of an operational amplifier is $\frac{R_f}{R_1} E_1$.

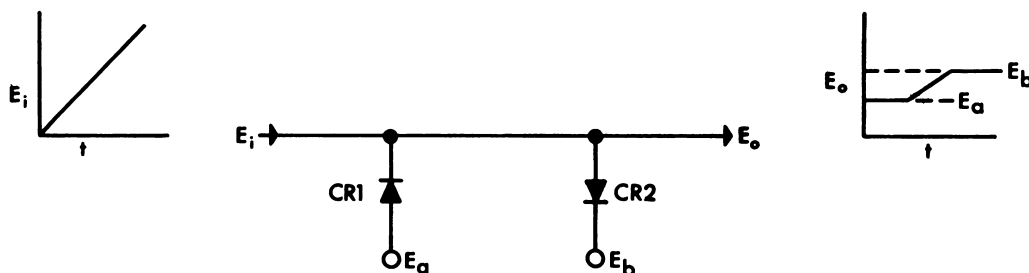
If R_1 is replaced by a servo driven potentiometer, the resistance of which is proportional to another voltage (say kE_2), the output of the operational amplifier will be $-\frac{R_f E_1}{kE_2}$. Thus, the output voltage is equal to the quotient, E_1 / E_2 , multiplied by a constant.



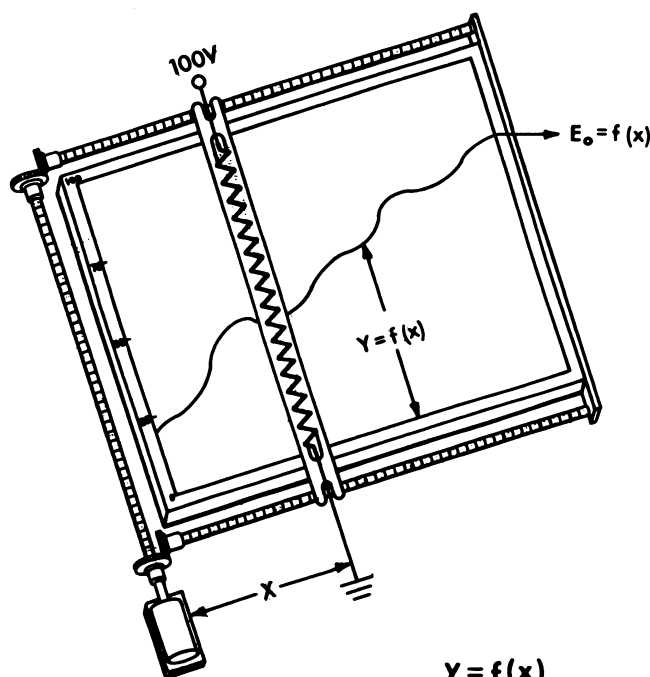
FUNCTION GENERATORS

A variety of devices are available to generate non-linear functions of a variable. These include the various types of cams described in Volume I. The most versatile electrical function generator uses diodes to break the function into segments. Arranged as shown, diode CR1 will conduct until $E_i = E_a$, and diode CR2 will begin to

conduct when $E_i = E_b$. The output voltage will be proportional to the input voltage only in the interval, $E_a \leq E_i \leq E_b$, and the slope can be controlled by an appropriate arrangement of resistors. By using a number of diode function generators, any function can be approximated by a series of straight line segments.

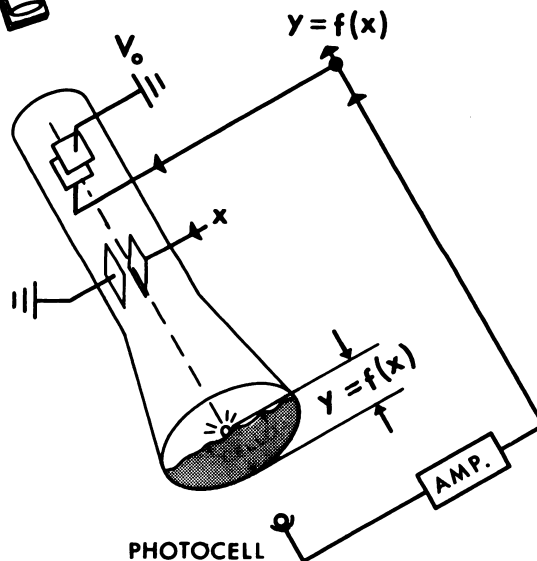


The servo table is an electromechanical device used to generate frequently used functions. A wire, the shape of which is determined by the function, is arranged in contact with a resistor. The resistor is then driven by a servo motor so that the displacement is equal to the independent variable (x). Since the point at which the wire meets the resistor is equal to $f(x)$, and the voltage at each point on the resistor is equal to y , the output voltage is equal to $f(x)$.



Where great speed or high frequency response is required, the fotoformer may be used as a function generator. With this device, a mask cut to the shape of the desired function is placed over the face of an oscilloscope. A photocell is placed in front of the scope face where it will respond to the intensity of the light spot.

The output of the photocell is connected to the vertical deflection plates, and an amplifier is adjusted so that a completely visible spot will create a signal large enough to drive the beam to the bottom of the scope face, and a completely masked spot will create sufficient voltage to drive the beam to the top. A state of equilibrium will occur when the spot is approximately half obscured by the mask. Therefore, when a voltage proportional to the independent variable (x) is applied to the horizontal deflection plates, the light spot will move along the top of the mask, and the vertical deflection voltage will be proportional to $f(x)$.



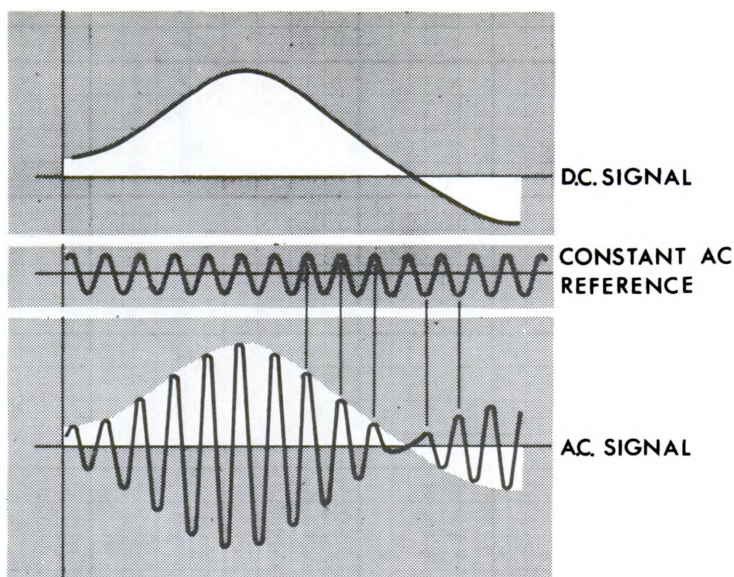
A.C. ANALOGS

It has been assumed so far that all electrical signals were direct current signals. Alternating current may also be used if the two following conditions are met: (1) - The frequency of the a.c. used must be greater than the maximum frequency response of the measuring devices used; (2) - If negative values of the variables are allowed, the devices used must be phase-sensitive.

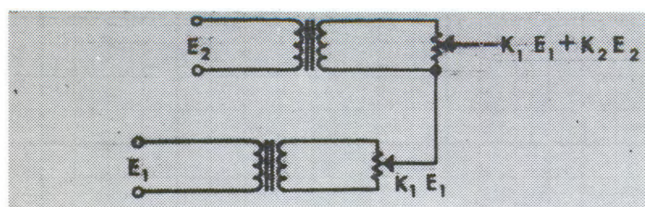
The illustration shows a d.c. signal and the same function represented by an a.c. voltage.

The instantaneous value of the a.c. signal does not indicate the value of the function, but the (r.m.s.) average value of the a.c. signal may be used to represent the value of a function. If the a.c. signal is the input to a servo motor, for example, the motor must not attempt to follow every variation of the a.c. signal, but must follow the average value. The second condition is essential because a negative a.c. signal does not exist. However, negative values can be indicated by a change in phase of the signal. Note that in the illustration, during the period when the d.c. signal is positive, the positive peaks of the a.c. signal correspond to the positive peaks of the a.c. reference, but during the period when the d.c. signal is negative the positive peaks of the a.c. signal correspond to the negative peaks of the reference; i.e., the signal is 180° out of phase with the reference.

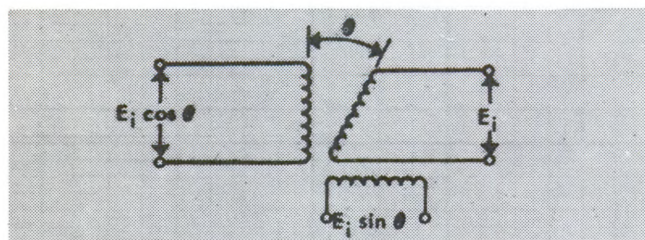
A.C. servo motors, which will rotate in one direction when the input signal is in phase with a reference voltage, and in the other direction when the signal is out of phase with the reference voltage are available. Potentiometer multipliers and servo table function generators will operate



equally well on a.c. or d.c. However, the operational amplifier, diode function generator, and fotoformer will not function with a.c. signals. The fotoformer, because of very high frequency response, will follow every variation in the signal instead of following the r.m.s. average value. The operational amplifier and diode function generator use tubes or semiconductors which conduct in one direction only, and therefore respond differently to the negative and positive portions of an a.c. signal. This disadvantage of using a.c. is offset by the fact that transformers can be used.



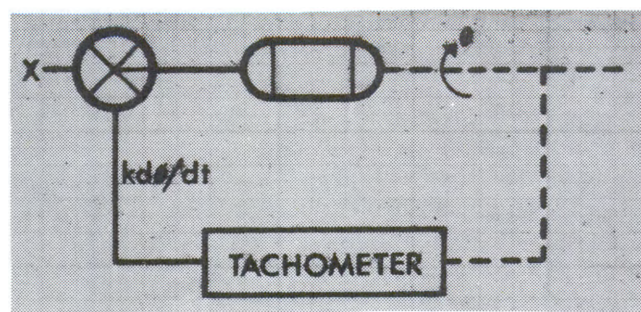
TRANSFORMER-COUPLED SUMMING AMPLIFIER



RESOLVER

A transformer-coupled summing network is illustrated. Three transformer windings, one of which is free to rotate, serve as a resolver.

The output of one side will be proportional to the input voltage times $\sin \theta$, and the output of the other side will be proportional to the input times $\cos \theta$.



SERVO-TACHOMETER INTEGRATOR

All electronic integrators for a.c. operation are also available, but are too complex for discussion here. The servo-tachometer integrator may be used with a.c. as well as d.c. The variable to be integrated is used to control a servo motor, which, in turn, drives a tachometer. The servo motor will rotate until the tachometer output ($k \frac{d\theta}{dt}$) is equal to the input voltage (x). Since $x = k \frac{d\theta}{dt}$, $\theta = \frac{1}{k} \int_0^t x dt + \theta_0$, where k is a constant of the tachometer, and the initial angle (θ_0) is the constant of integration.

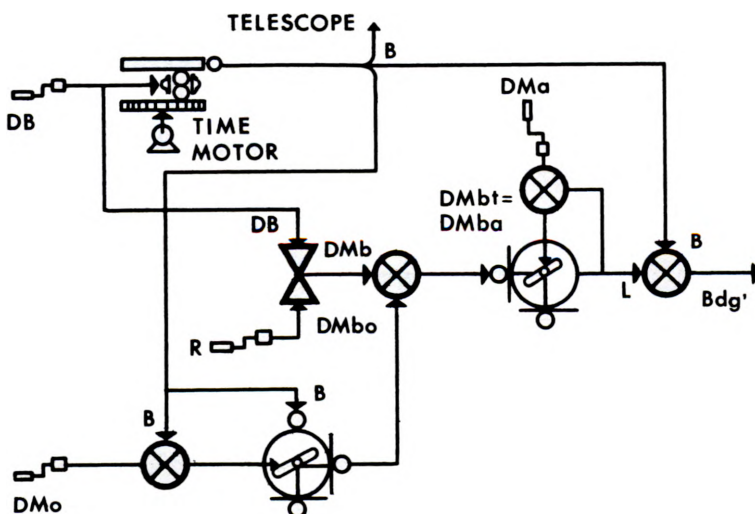
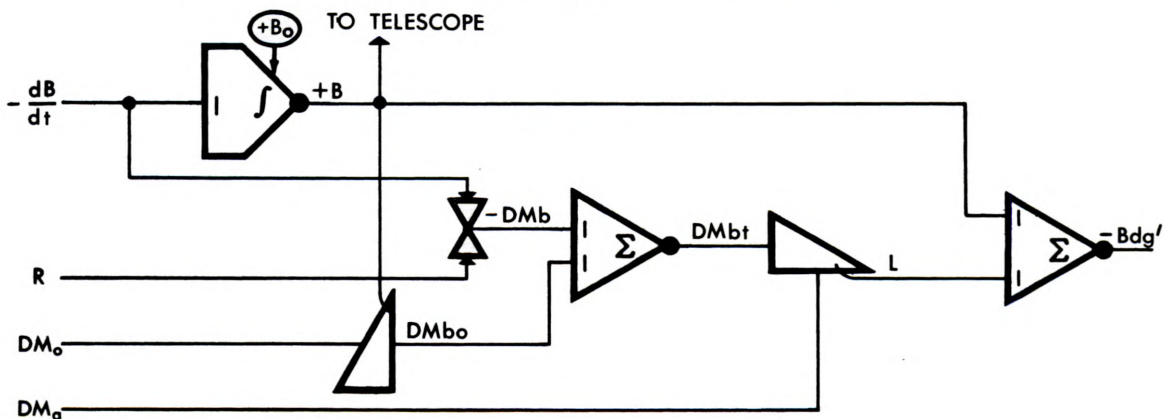
ROAD MAPPING

Before discussing the applications of the various devices just described, a brief discussion of the symbols used to diagram analog computers is in order. When mechanical computing devices were discussed in Volume I, symbols which pictorialized the devices most commonly used to perform the function were used. With electronic and electromechanical computers, there are too many devices used for most functions for a representative device to be selected, and the electronic diagram may be of no help to someone unfamiliar with electronic circuitry. For electronic and electromechanical devices, there-

fore, purely arbitrary symbols are used, except in the case of differentials used with servomotors. Since the differential is commonly considered part of the servo-mechanism, the same symbol is used for an electrical and mechanical differential.

The process of diagramming a computer arrangement to solve a given problem is called road mapping. There is no difference in the techniques used for electromechanical computers, and the methods employed in Volume I for mechanical computers.

MECHANICAL SYMBOLS						
ELECTRIC SYMBOLS						



Compare the diagram for a mechanical analog computer to solve the torpedo tube train problem, reprinted from Volume I, with the road map for an electronic computer to solve the same problem. The electronic computer eliminates two compensating differentials, which are not required with electronic resolvers. With this exception, there is a one-to-one correspondence between elements of the mechanical and electronic computers. (No provision for setting the initial value of bearing (B) was mentioned. It is present, however, since this setting will be forced on the integrator by the telescope operator.) There is no essential difference between the road map for an electronic computer, and that for a mechanical computer; the only difference is in the construction of the devices used. The procedure used to mechanize problems (which is the next topic) applies to electronic and mechanical computers.

APPLICATIONS OF ANALOG COMPUTERS

The uses of analog computers fall into three general categories -- solution of equations, simulation, and control. The difference is more in the method by which the analog is arrived at than in the actual nature of the analog. For example, if analog devices were used to simulate a mechanical system piece by piece, it would be called simulation. If the differential equations for the system were obtained, and analog devices used to solve the equations, it would be called solution of equations, yet the analogs resulting from both processes might be identical.

SOLUTION OF EQUATIONS •

As an application encountered in weapons systems, consider the problem of generating range and bearing for a tracking system. These quantities are normally measured, but the tracking may be interrupted, and it is necessary to coast. Assuming the target continues to move

with the same velocity as at the time tracking is interrupted, the range (R) and bearing (B) can be generated from the components of velocity along the LOS (\dot{R}) and tangential to the LOS ($R\dot{B}$). (The dot indicates a derivative with respect to time. Thus, $\dot{R} = \frac{dR}{dt}$, and $\dot{B} = \frac{dB}{dt}$.)

The illustration shows a target moving on a straight line from P_1 to P_2 with velocity V . As the bearing changes from B_1 to B_2 , the angle between the velocity vector (V) and the radial component of velocity (R) changes from θ_1 to θ_2 . Since lines OP_1 and XP_2 are each at an angle θ_1 to the velocity vector, XP_2 is parallel to OP_1 and $\Delta\theta = \Delta B$. Since $\Delta\theta = \Delta B$, $d\theta/dt = dB/dt = \dot{B}$.

Resolving the velocity vector into radial and tangential components gives:

$$\begin{aligned}\dot{R} &= V \cos \theta \\ R\dot{B} &= V \sin \theta\end{aligned}\tag{1} \tag{2}$$

The time rate of change of radial velocity is

$$\frac{d(\dot{R})}{dt} = -V \sin \theta \frac{d\theta}{dt}\tag{3}$$

Substituting equation (2) into equation (3)

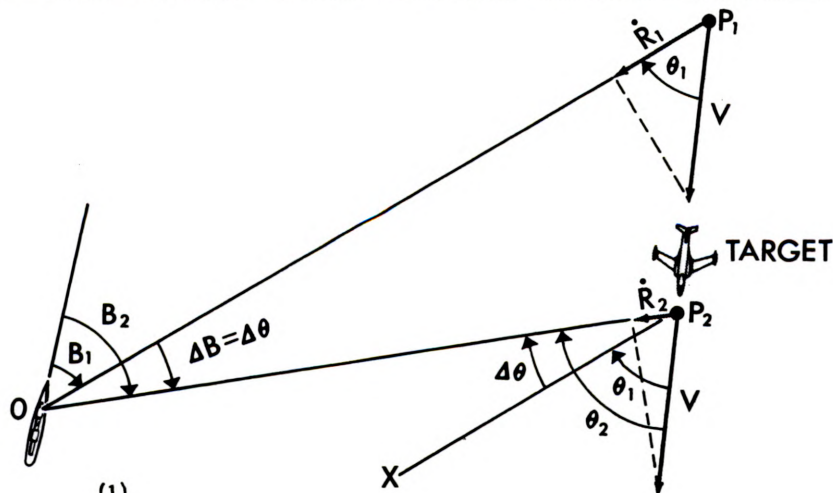
$$\frac{d(\dot{R})}{dt} = -R\dot{B} \frac{d\theta}{dt}\tag{4}$$

and since $\frac{d\theta}{dt} = \frac{dB}{dt} = \dot{B}$

$$\frac{d(\dot{R})}{dt} = -R(\dot{B}^2)\tag{5}$$

and $\dot{R} = -\int_0^t R(\dot{B}^2)dt + (\dot{R})_0$

where $(\dot{R})_0$ is the initial radial component of velocity.



From equation (2) the time rate of change of tangential velocity is

$$\frac{d(R\dot{B})}{dt} = V \cos \theta \frac{d\theta}{dt}\tag{7}$$

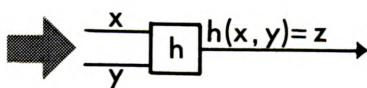
Again using the fact that $\frac{d\theta}{dt} = \frac{dB}{dt} = \dot{B}$, and substituting equation (1) into equation (7)

$$\frac{d(R\dot{B})}{dt} = \dot{R} \frac{d\theta}{dt} = \dot{R}\dot{B}\tag{8}$$

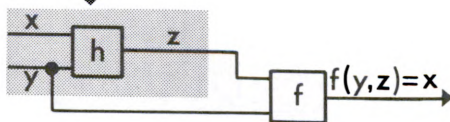
and $R\dot{B} = \int_0^t R\dot{B} dt + (R\dot{B})_0$

where $(R\dot{B})_0$ is the initial tangential component of velocity.

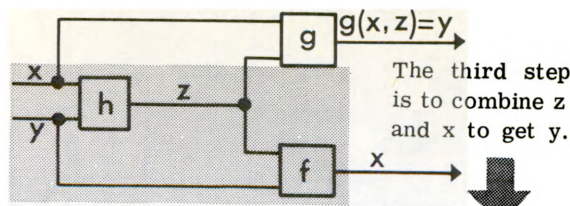
The solution of equations requires a process of circular reasoning. One or more variables are assumed to be available, and are used to form other variables, until the variables originally assumed are obtained as an output. For example, given three equations, $x=f(y,z)$, $y=g(x,z)$, and $z=h(x,y)$, any two variables must be chosen as a starting point.



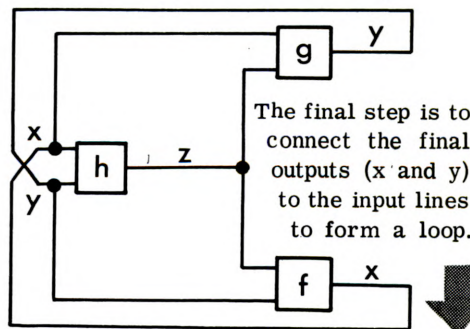
Starting with x and y , the first step is to combine x and y to get z .



The second step is to combine y and z to get x .



The third step is to combine z and x to get y .

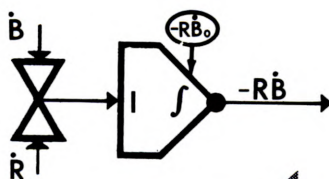


The final step is to connect the final outputs (x and y) to the input lines to form a loop.

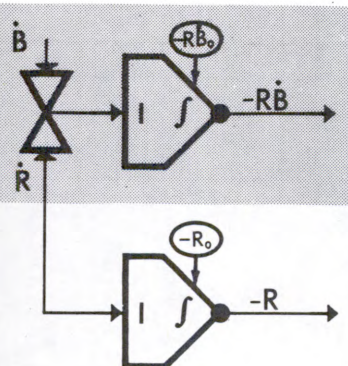
The result will be the same no matter which variables are chosen as the starting point.

The equations to be solved (see box) are:

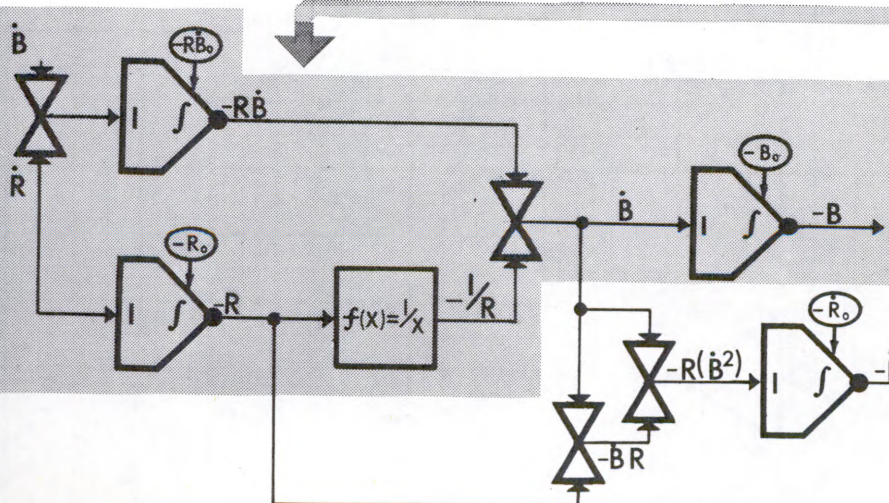
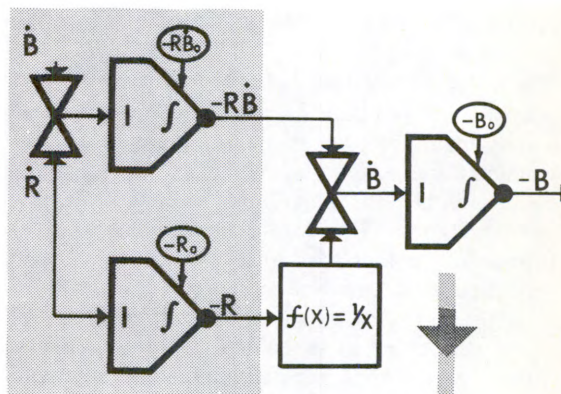
$$\begin{aligned} R &= R_0 + \int_0^t \dot{R} dt \\ B &= B_0 + \int_0^t \dot{B} dt \\ \dot{R}(t) &= -\int_0^t R (\dot{B}^2) dt + (\dot{R})_0 \\ R\dot{B}(t) &= \int_0^t \dot{R} \dot{B} dt + (R\dot{B})_0 \end{aligned}$$



The first step toward the road map is to select some variable as a starting point. Beginning with \dot{R} and \dot{B} , combining them according to the fourth equation, yields $-R\dot{B}$, and requires one multiplier and one integrator.

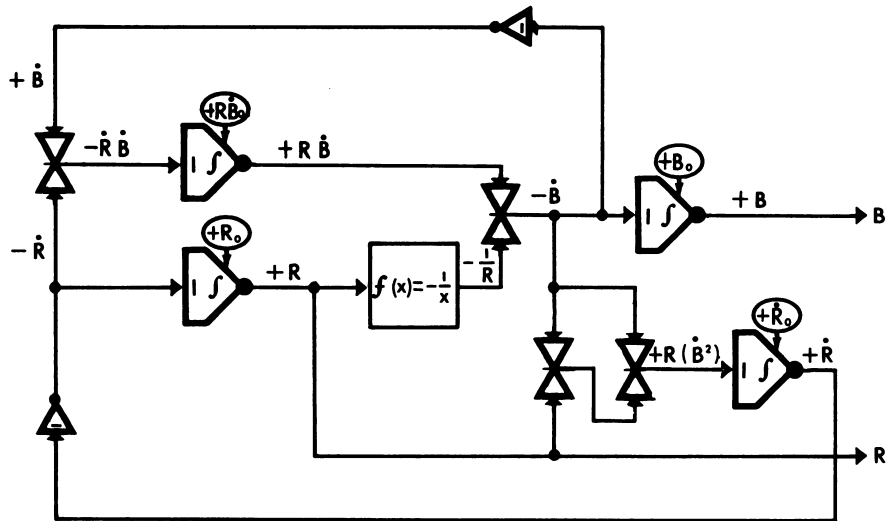


This gives the first part of the road map. Since \dot{R} is available, it is integrated according to the first equation, to yield $-R$ as the second step in road mapping the solution. Since $-R\dot{B}$ and $-R$ are now available, the product $\frac{1}{R} \times R\dot{B}$ will yield \dot{B} , which in turn may be integrated according to the second equation, to yield B as the third step toward the solution.



The only equation remaining is the third, so two multipliers are used to generate $-R(\dot{B}^2)$ from $-R$ and \dot{B} , and the result is integrated to yield $-\dot{R}$. Both the variables which were assumed at the beginning (\dot{R} and \dot{B}) are seen to be available, but in negative form in the case of \dot{R} . All that remains is to connect these points to the starting points, and bring out the lines for R and B . Three operational amplifiers could be used to perform the conversion, but by

juggling signs throughout the road map, the result can be achieved with only two, as has been done in the final illustration. (One inversion has been accomplished by using a function generator for $-1/x$ instead of $+1/x$.) This gives an additional advantage in that positive initial values instead of negative valves for range, bearing, and velocity are set into the integrators.



The solution of the equation $A\ddot{x} + B\dot{x} + Cx + D = 0$ illustrates the typical procedure for solving differential equations. Any one of the three variable (x , \dot{x} , \ddot{x}) could be chosen as the starting point. If either x or \dot{x} is chosen, however, at least one differentiation is necessary. As pointed out before, differentiators are to be avoided whenever possible, so \ddot{x} should be chosen as the starting point. Successive integration gives \dot{x} and x . Multiplying these by the appropriate constants, and combining them according to the equation, $\ddot{x} = -\frac{B}{A}\dot{x} - \frac{C}{A}x - \frac{D}{A}$,

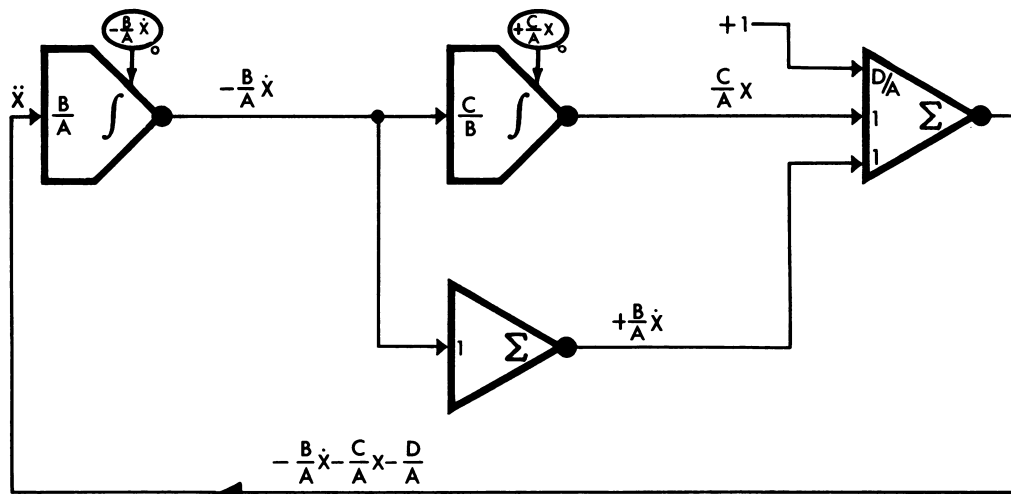
gives \ddot{x} , which is then connected to the first operator, closing the loop.

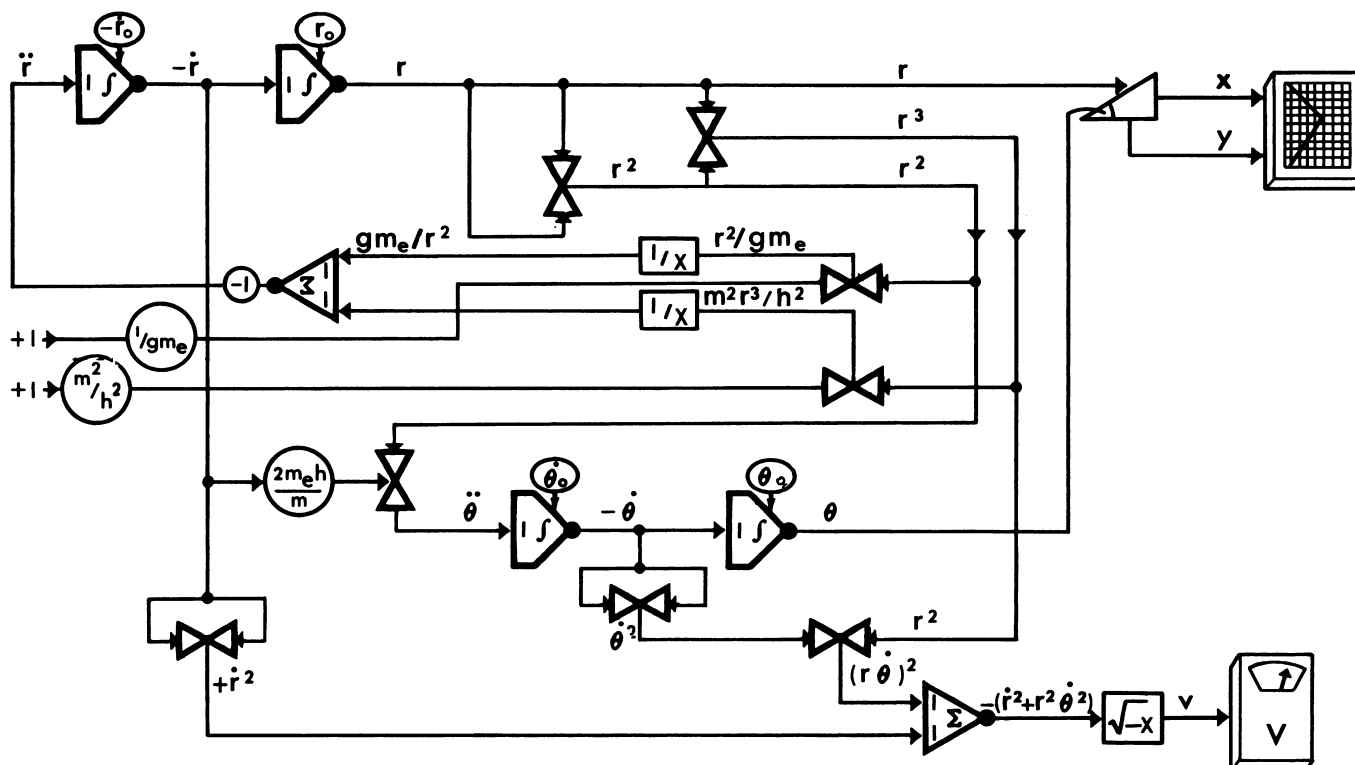
The initial conditions, i.e., the values of \dot{x} and x at $t = 0$, must be set into the integrators. These values are usually available, but there is a class of equation, called boundary-value, or eigenvalue problems, where other conditions at later times must be met, (i.e., the problem is to determine what initial conditions will give a certain result, rather than to determine the result of given initial conditions). For the solution of this type of equation, it is necessary to experiment with different initial values until a combination which meets the specified conditions is found. A repetitive type analog computer is normally used for this type of problem. The non-re-

petitive or one-shot computers described so far will start at $t=0$ and continue running until stopped. A repetitive type will run for a predetermined time, and then return to the situation for $t=0$ repeating the same process several times a second. As long as the settings remain unchanged, the same result will be repeated. The value of this type computer is that the operator can vary the initial condition settings slightly, and continuously observe the effect of these changes on the result. Assume, for example, the value of the initial conditions is desired, which meets specified conditions at a later time. Such a problem might arise in a ballistic missile trajectory calculation to launch a satellite, where the operator knows the final range, bearing, and velocity which will result in the desired orbit, and wishes to learn what initial (burnout) conditions are necessary to achieve this result, and the amount of time the flight will take (i.e., at what time the satellite should be released). The equations have been derived in the chapter on missile flight paths, (Chapter 4) and may be written:

$$r = \frac{Gm_e}{r^2} + \frac{h^2}{m^2 r}, \quad \theta = \frac{-2m_e h}{mr^2}, \quad V = \sqrt{\dot{r}^2 + r^2 \dot{\theta}^2}$$

h = angular momentum = constant for each burnout situation.





The road map for a computer arrangement to solve these equations is shown. The points corresponding to $r \sin \theta$ and $r \cos \theta$ would be connected to an x-y plotter, or to an oscilloscope. The values of r , θ and v would also be recorded. The operator could then select the appropriate initial conditions, for R , \dot{R} , θ , $\dot{\theta}$, and h . (In actual practice, since $h = m r_0^2 \theta_0^2$, connections would be included to set the potentiometers for m^2/h^2 and $2m_e h/m$ automatically as r_0 and θ_0 are set.) As the operator changes the initial condition each time the computer recycles, he can observe the effect on the flight and path and velocities, as pictured on the recorders. It is then a simple matter to find the initial conditions which give the desired results. Since the computer is operating on real time, i.e., 1 second of computer time equals 1 second of flight time, the flight time is the interval between the start of a computer cycle and the achievement of the desired range, bearing, and velocity. Most analog computer devices integrate or differentiate only with respect to time. This is not a serious limitation, since time can be substituted for the independent variable in functions which do not involve time. Assuming a problem involving $\int x dy$ and $\int y dx$, t can be substituted for x , giving $\int t dy$ and $\int y dt$. Furthermore, $\int t dy = yt - \int y dt$, so the problem is reduced to integrals with respect to time only.

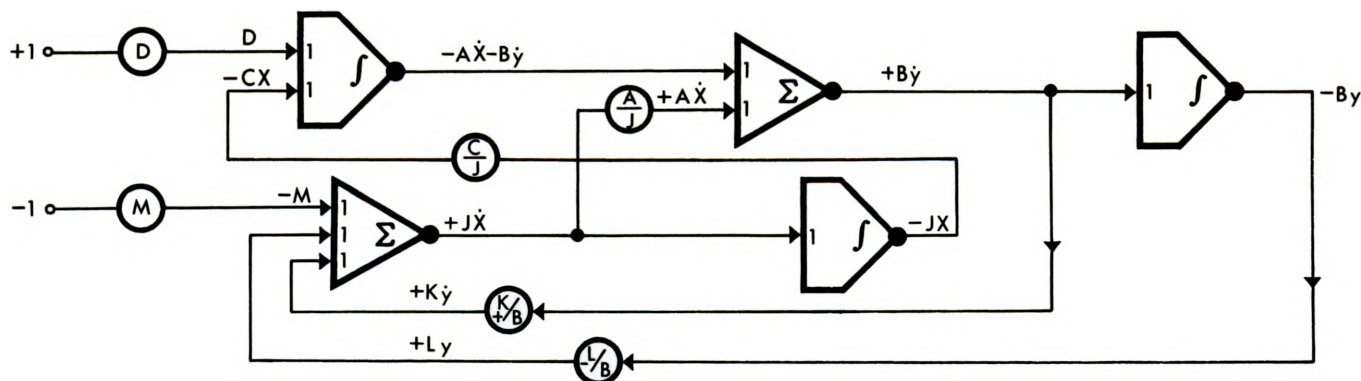
IMPLICIT FUNCTIONS

When the devices to solve an equation directly are not available, the computer may be programmed to solve an implied equation. If the equation to be mechanized is $\sqrt{y} = x$, and a square root circuit is not available, the implied equation $y - x^2 = 0$ may be solved instead.

Similarly, given the simultaneous differential equations: $A\ddot{x} + B\ddot{y} + Cx = D$, and $J\dot{x} + K\dot{y} + Ly = M$, the standard procedure calls for solving one equation for \ddot{x} and the other for \ddot{y} . This is not possible in this case, since the first equation can be solved for only one of these quantities, and the second equation cannot yield either.

This problem may be circumvented by solving the implied equation $J\ddot{x} + K\ddot{y} + L\dot{y} = 0$, obtained by differentiating the second simultaneous equation.

Unfortunately, these techniques usually introduce false solutions, or reduce the number of solutions. The implied equation may have more than one solution, some of which are solutions of the original equations, and some of which are not, or some (but not all) are solutions of the implied equation. The first example, has two real solutions for each value of x , the implied equation has only one real solution for each value of x . In the second example, since M has been eliminated from the implied equations, the solutions are independent of M . The solution of the original equation is not independent of M , so the original equation must have fewer solutions than the implied equation. A better approach to the solution of the simultaneous differential equation is to solve for $A\ddot{x} + B\ddot{y}$. Since $D - Cx = A\ddot{x} + B\ddot{y}$ from the first equation; $\int (D - Cx) dt = A\dot{x} + B\dot{y}$. The second equation can be solved for \dot{x} , and $A\dot{x}$ subtracted from $A\dot{x} + B\dot{y}$ to give $B\dot{y}$. Having both \dot{x} and \dot{y} the solution is easily obtained without introducing false solutions. As an exercise in these methods of programming, it is recommended that the reader draw a road map for the solution of these simultaneous differential equations before looking at the solution given.

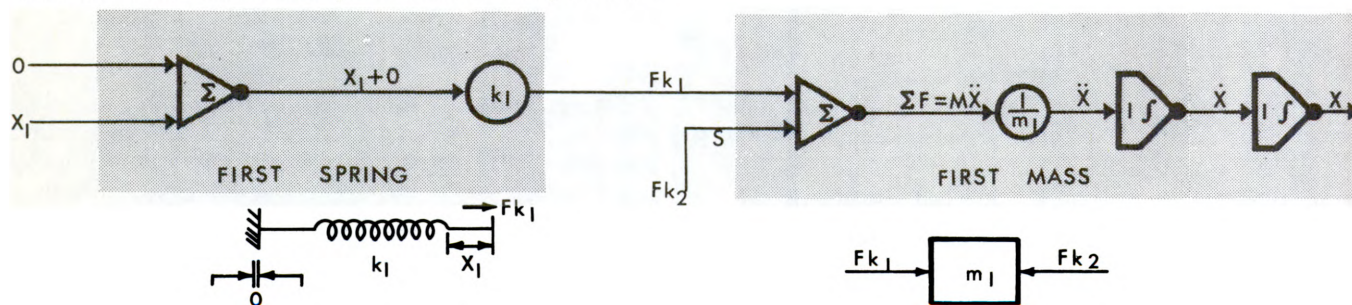


ROAD MAP FOR SOLUTION OF
 $Ax + By + cx = D$
 $Jx + Ky + Ly = M$

SIMULATION

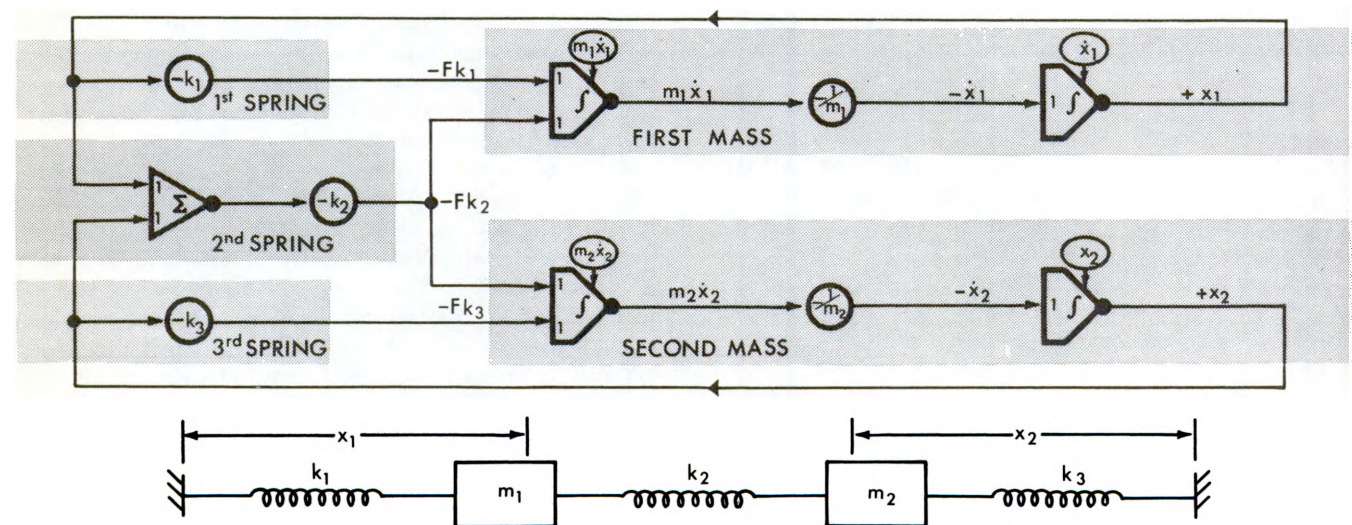
The use of passive analogs to simulate mechanical systems was discussed earlier where a passive analog of two masses on three springs was constructed. The same system can also be simulated by an active analog. Each mass is simulated by a summing amplifier, which sums the forces on the mass, and since $F = m\ddot{x}$, the sum

of the forces is multiplied by $1/m$, and integrated twice to give the displacement. Each spring is simulated by a potentiometer which multiplies the total displacement by the spring constant to give the force. No attention is given to signs, or to superfluous amplifiers.



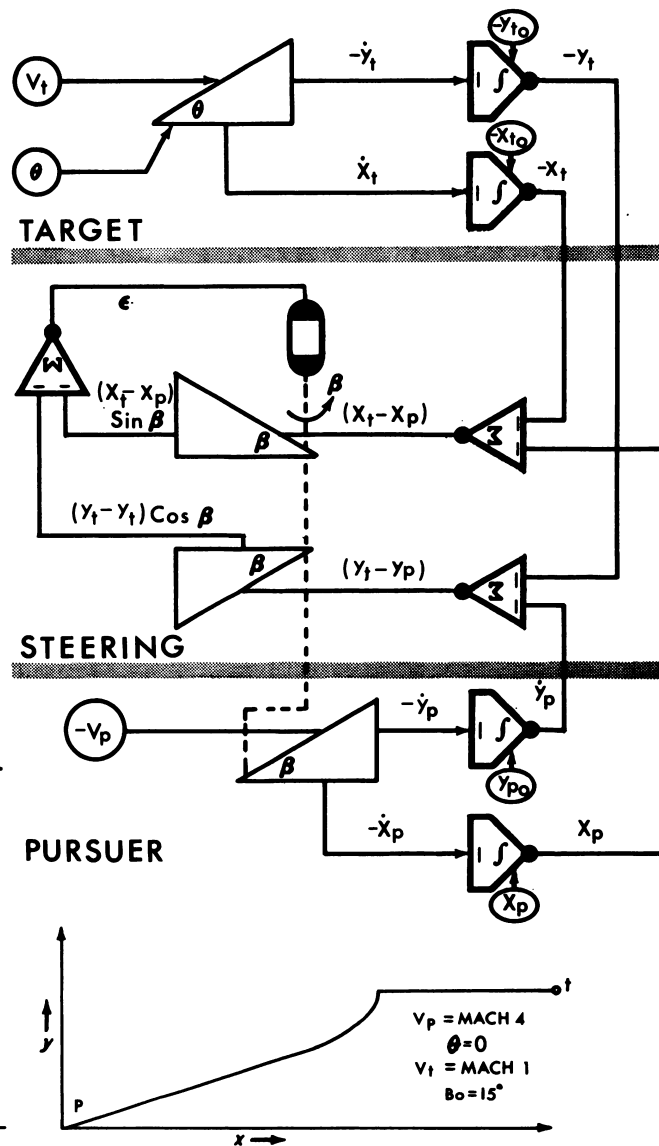
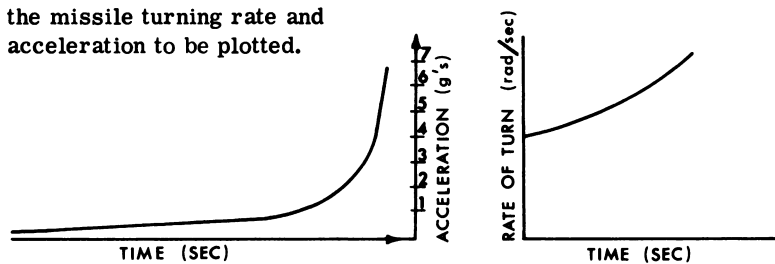
The complete analog is shown with the correct signs, and without superfluous amplifiers. Comparing this analog to the passive analog described earlier, reveals the typically large increase in the number of components. As a more complex problem, consider the pursuit course

simulation. Here, the problem is to simulate a pursuer (airplane, guided missile, ship, etc.) and a target, with the provision that the pursuer always heads directly toward the target.



The resolver at the upper left yields the x and y components of the target velocity. Integration of these parameters gives the target position components (x_t , y_t) from which the corresponding components of the pursuer are then subtracted to give the horizontal and vertical components of the distance to the target. Note that these values ($x_t - x_p$) and ($y_t - y_p$) form two sides of a triangle, and that $(y_t - y_p) \div (x_t - x_p) = \tan \beta = \sin \beta \div \cos \beta$. When the proper value for β is reached, $(y_t - y_p) \cos \beta = (x_t - x_p) \sin \beta$, and the input to the servo motor (ϵ) will be zero. If β deviates from this value the servo will be driven until the geometry is again satisfied. This servo motor also drives a resolver which yields the x and y components of the pursuer's velocity. When these terms are integrated, they yield the position components of the pursuer.

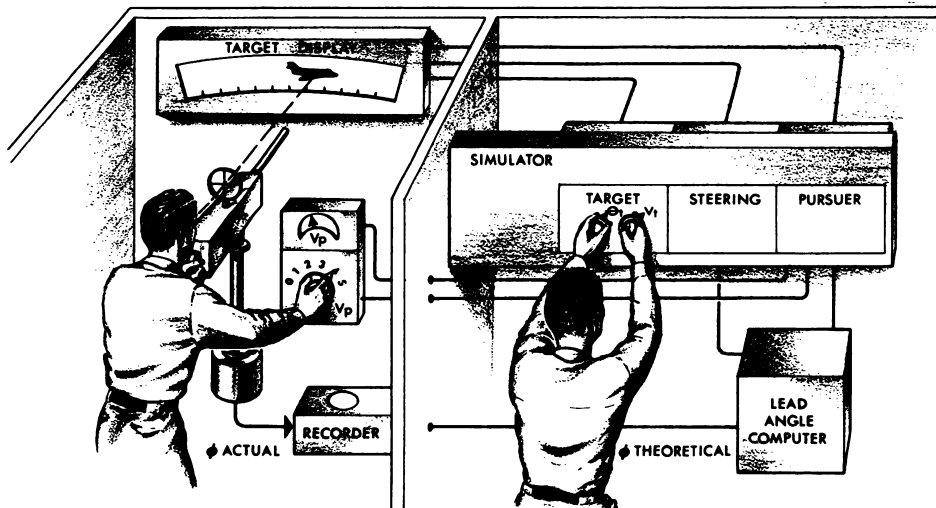
This simulator could be used to study the trajectories of a homing missile. By connecting the x_p and y_p outputs to a recorder, the course of the missile can be plotted for any desired situation. Since the turning rate of the missile (β) is equal to the turning rate of the servo motor, the addition of suitable instrumentation will also enable the missile turning rate and acceleration to be plotted.



Another interesting application of this computer would occur in testing an aircraft gunsight with a human operator. The target maneuvers would be created by an operator manipulating the potentiometers controlling

the target heading (θ) and velocity (V_t). The target bearing (β) and velocity would be used to create a target display for the gunner. The pursuer velocity is provided for the gunner (a gunner would know the

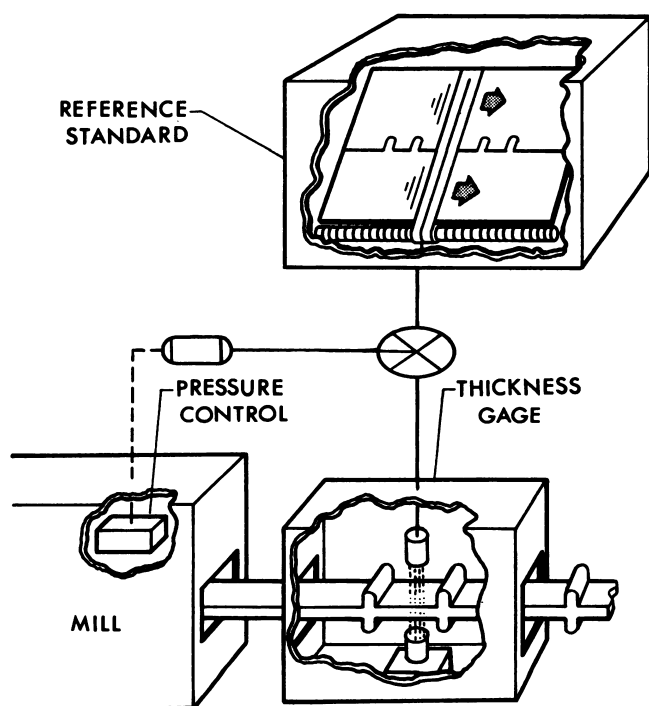
velocity of his own craft), and he must estimate the bearing and velocity of the target, and using the real gunsight, aim and "fire" at the target. The actual heading of the "bullet" (ϕ) will be recorded, while at the same time the true simulated target bearing and velocity are used by another computer to determine the theoretical best heading of the bullet, which is also recorded. Comparing the actual value of ϕ with the theoretically perfect value will reveal the worth of the gunsight, or the operator - gun sight system.



CONTROL

In addition to performing calculations, analog devices are frequently used to translate the results into action. The most common application of this type is in the area of weapon control and system control, and is covered in detail in other chapters. Another application can be classified as "continuous process control" in which analog devices monitor a process and make adjustments so the output conforms to some standard or variable criterion. Such control systems are used extensively in industry, particularly in fabricating, milling and refining processes. Analog devices, based on the "pantograph" principle, are used to control machine tools in order to turn out parts matching a model or a drawing. More complex systems are used in petroleum refining processes, where they adjust the various cracking and distilling operations to produce varying proportions of different end products in order to meet varying demands.

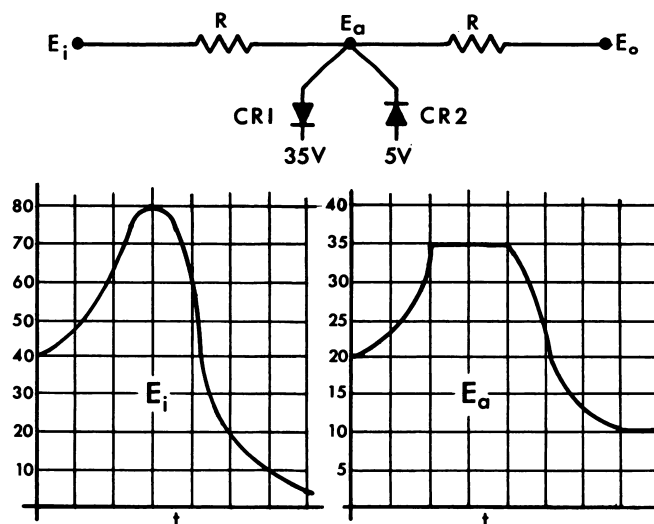
Although every control system is usually unique, each can be categorized somewhat in terms of the parameter monitored. Many systems monitor some physical characteristics of the output, such as the thickness of sheet metal from a rolling mill. The measured thickness is compared with the desired thickness and the roller pressure adjusted to correct any difference. It is not necessary that the desired thickness be uniform. If the reference standard is variable, the output will vary also. Other systems monitor the demand and adjust the output rate to conform. Almost all large electrical power distribution is controlled according to demand. More complex systems monitor several aspects of the process and strive for maximum efficiency of operation or for maximum profit.



LIMITS AND STABILITY

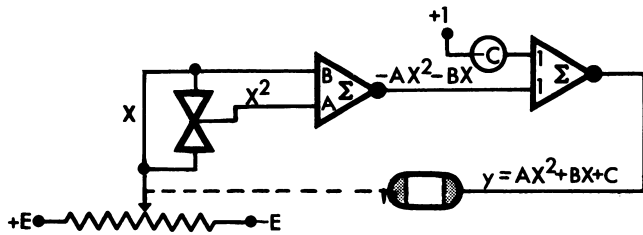
Two aspects of analog computers have been largely ignored in the previous discussions: limits and stability. The previous application of analogs in the simulation of aircraft for testing a gunsight illustrates the need for one type of limit in analog devices. If the computer just discussed is to provide a useful simulation, the various parameters must be constrained to reasonable values. Since a real aircraft has limits on its speed and maneuverability, the simulator, to be realistic, must also have limits.

Assume the target aircraft has a maximum velocity of 700 knots, and will stall below 100 knots. Assume further that 1 volt in the computer corresponds to 10 knots. The input voltage to the resolver must be constrained to the region from 10 to 70 volts, i.e., $100 \leq V_t \leq 700$. The limits are usually imposed by placing diodes in the circuitry, as shown. Since the voltage at the junction is one-half the input voltage, while the input voltage remains between 10 and 70 volts, neither diode will conduct, and the output will be unaffected by the presence of the diodes. However, if the voltage rises above 70, CR1 will conduct, the voltage at the junction will remain at 35 volts regardless of the value of the input voltage, and the output will be the same as when the input was 70 volts. The effect of CR2 is similar at the minimum voltage. A similar arrangement on the input signal to the servo (ϵ) will limit the pursuer maneuverability ($\frac{d\beta}{dt}$).



The second need for limits is created by the nature of active analog devices. In all previous computer arrangements, the solution has been dynamic; i.e., the solution sought was a curve or range of permissible values of the variables. In some cases, the solution required is a specific number. Such a problem is that of finding the roots of a quadratic equation (more generally: given $y = Ax^2 + Bx + C$, find the values of x for which $y=0$). Theoretically, the computer arrange-

SCALE FACTORING



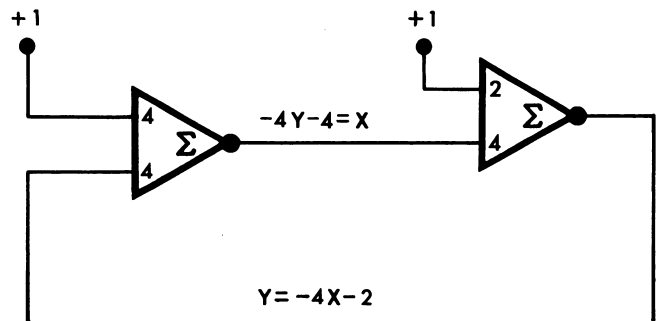
ment shown will cause the servomechanism to hunt until a solution is found. In practice, there are problems. Assume the servomechanism is such that a positive input ($y > 0$) causes the potentiometer to move

toward a lower value of x , and vice versa, and the specific equation is $y = x^2 + 3x + 2$. The graph shows that there are two solutions to the problem: $x = -1$ and $x = -2$. If the potentiometer wiper is initially in the region where $x > -1$ and $y > 0$, and the wiper will move toward $x = -1$. If the initial position is between $x = -2$ and $x = -1$, y is negative, so the servo will again move toward $x = -1$. Therefore, as long as the initial value of x is greater than -2 , the computer will arrive at the solution, $x = -1$, and equilibrium will occur at this point. If the potentiometer wiper is initially at a point where x is less than -2 , i.e., a point to the left of $x = -2$, y is positive, and the servomechanism will move toward lower values of x and continue indefinitely without reaching a solution. If the potentiometer wiper is initially at $x = -2$ exactly, equilibrium will exist, and the computer should remain at this point. The equilibrium is unstable, however, and any disturbance will dislodge it. This loop is said to be unstable because of positive feedback. Note that a value of x which is too low after passing through the series of amplifiers, will create a new value which is still lower. It is also possible to attempt to solve an equation for which no real solution exists; for example, $x^2 + 3x + 3 = 0$, in which case the computer would run indefinitely. Another problem might occur if the value of $\pm E$ on the potentiometer is too small to allow the proper value of x to be reached. One

method of providing for these contingencies is to put limit stops on the servomechanisms. Mechanisms to accomplish this were described in Volume 1. In this particular case, a limit stop on the potentiometer wiper at both extremes which will stop the servo, and then reverse the input connection to the servo, would solve the problem of the computer running away from, rather than toward, the solution. Further circuitry, which would stop the computer if both limits were reached and provide an indication that this had occurred, would solve the problem of non-existent solutions. The problem of insuring a sufficient voltage range on the potentiometer is handled by scale factoring.

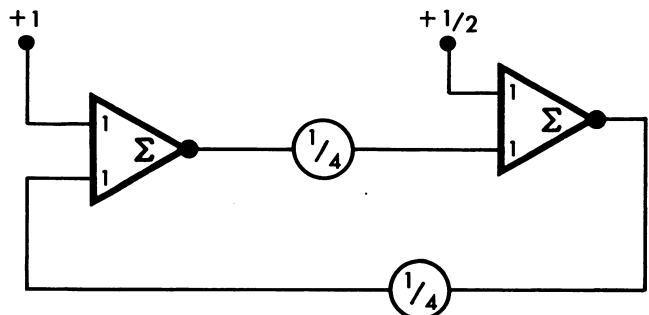
method of providing for these contingencies is to put limit stops on the servomechanisms. Mechanisms to accomplish this were described in Volume 1. In this particular case, a limit stop on the potentiometer wiper at both extremes which will stop the servo, and then reverse the input connection to the servo, would solve the problem of the computer running away from, rather than toward, the solution. Further circuitry, which would stop the computer if both limits were reached and provide an indication that this had occurred, would solve the problem of non-existent solutions. The problem of insuring a sufficient voltage range on the potentiometer is handled by scale factoring.

To use an analog it is necessary to establish a scale factor on each quantity, i.e., to determine how many volts or how many degrees of rotation will equal one foot, or one pound, etc. A scale factor for time is also required. In the previous examples we have considered all scale factors equal to one. When the scale factor for time is one, the result is called a real time analog, since the analog will complete a process in the same time as the subject. This is not always desirable, since the conditions studied may occur too rapidly for proper evaluation, or so slowly that an unreasonable amount of computer time is required. The ability to slow down or speed up an operation by scale factoring time is one of the most useful features of analog computers. In the case of quantities other than time, unit scale factors are usually desirable, but not always practical. Consider the solution of the simultaneous equations, $4x + y + 2 = 0$ and $x + 4y + 4 = 0$. In the road map shown, the loop has positive feedback with a

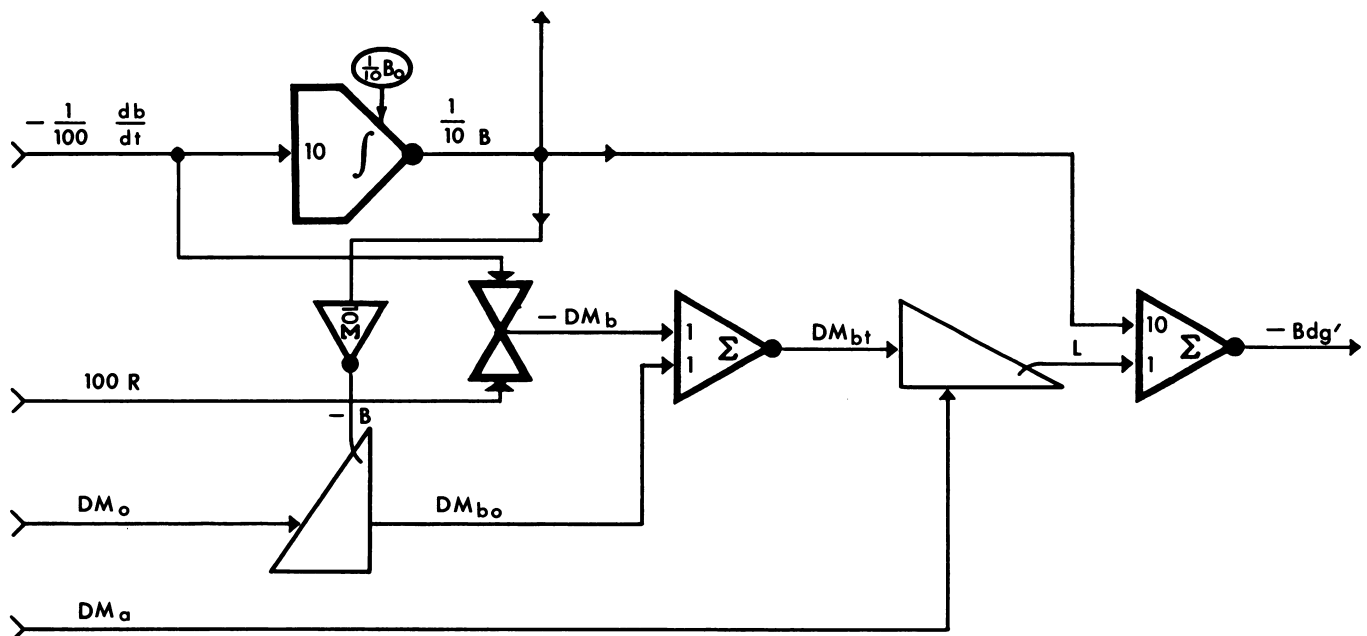


gain of 16, and therefore is unstable. As in the solution of the quadratic equation discussed before, if the initial values are exactly at the solution ($x = -4/15$, $y = -14/15$) the loop will be in equilibrium, but any deviation will create a runaway condition. Many techniques are available to correct for this condition.

In this instance, dividing each equation by four is sufficient. This gives $x + 0.25y + 0.5 = 0$, and $0.25x + y + 1 = 0$, and the road map shows that the loop is stable.



A more complex problem is scaling the road map for the torpedo tube train computer presented earlier. The road map is reprinted here with scale factors added.



The first step is to assume maximum and minimum values for the variables. Assume a range of 500 to 5,000 yards, and maximum speeds for both ships of 20 yd/sec (approximately 40 knots). The bearing can range from -2π to $+2\pi$ radians. (Limits of $-\pi$ and $+\pi$ cannot be used, since the variation must be continuous, and cannot suddenly switch from positive to negative values as the bearing passes 180° .) The maximum value of \dot{B} will occur when the target is moving tangentially to the LOS at the minimum range and at maximum speed. Therefore, $\dot{B}_{\max} = 20 \text{ yd/sec} \times 500 \text{ yd} = 0.04 \text{ rad/sec}$. Since the computer value of \dot{B} is integrated to drive the telescope, the value of \dot{B} must be allowed to exceed the actual target bearing rate for brief intervals in order to permit the telescope to catch up with the target. Therefore, the limits of \dot{B} will be taken as $\pm 0.15 \text{ rad/sec}$. The torpedo velocity is taken as 40 yd/sec (about 75 knots).

A typical operational amplifier operates with an upper voltage limit of 100 volts. There is no lower limit, but, in order to maintain accuracy, it is desirable to maintain the voltage as high as possible (within the 100-volt limitation). As a starting point, choose factors for \dot{B} and B : since $-0.15 \text{ rad/sec} \leq \dot{B} \leq +0.15 \text{ rad/sec}$, 1 volt = 0.01 rad/sec is convenient for \dot{B} , and since $(-2\pi \text{ rad} \leq B \leq +2\pi \text{ rad})$, 1 volt = 0.1 rad is convenient for B .

These values will limit the voltage for \dot{B} to the range from -15 to +15 volts; and for B to the range from -63 to +63 volts. A convenient factor for range is 1 volt = 100 yards. Since DM_b is the product of R and \dot{B} , the scale factor is the product of the scale factors of R and \dot{B} , or 100 yards \times 0.1 rad/sec = 1 yd/sec per volt. Since the maximum value of DM_b is 20 yd/sec, this is a reasonable factor. Since the factor for B is

.1 rad/volt, and the factor for \dot{B} is .01 rad/sec/volt, a gain of 10 is required of the integrator. For convenience, the factors for DM_O and DM_a are chosen equal to the factor for DM_b . Since a factor of .1 is being used for B , the final amplifier must multiply B by 10 to get the proper value for Bdg' . The remaining step is to convert $0.1B$ to B for the final resolver.

This is accomplished by the addition of an amplifier not contained in the original road map. (The output is $-B$ instead of B , but since $\cos(-B) = \cos B$, it makes no difference.)

As a problem becomes more complex, many compromises must be made, because the scale factor chosen for one quantity affects later quantities which are functions of the first quantity scaled. However, the principles remain the same no matter how complex the road map.

None of the illustrative problems employed a scale factor for time. The ability to use this factor is one of the most useful characteristics of an analog computer. It permits fast operations to be slowed down and studied at leisure, and slow operations to be simulated quickly. Note that the only devices in which time appears as a variable are the integrators and differentiators. The mechanical integrators represent time by a mechanical rotation induced by a constant speed motor. The computer time can be varied by speeding up or slowing down the motors. This is not possible with other types of devices, since the operation of the device is a fixed function of time. However, since the input to any integrator is equal to the time rate of change of the output, the computer can be made faster than real time by multiplying the gain of each integrator by the desired factor (or by dividing, if a computer time slower than real time were desired).

PROGRAMMING

The process of stating a problem in terms which can be handled by a computer, and of determining the arrangement of elements which will solve the problem, is called programming. The methods used to derive and scale the road maps for the examples given throughout the chapters are typical of the procedure.

There is a great difference between the programming of a special purpose and a general purpose computer. The special purpose computer is programmed only once, before it is built. Since each component is used for only one purpose, there is little reason to build a device with a capability greater than required. The special purpose computer, which is the type most frequently found in weapons systems, often contains devices with limited versatility, but which are efficient at the specific calculation for which they are designed.

The general purpose computer is programmed as each problem occurs, and will be programmed many times during its lifetime. Components which are versatile and easily programmed are essential.

After the program for an analog computer is conceived, it must be checked for stability, adequate response time, sufficient range, and for the existence of solutions. This last factor can cause considerable difficulty.

A computer might be designed to solve the quadratic equation $x^2 + 3x + 3 = 0$, for example. There is no non-imaginary solution to this equation, and the computer would run indefinitely.

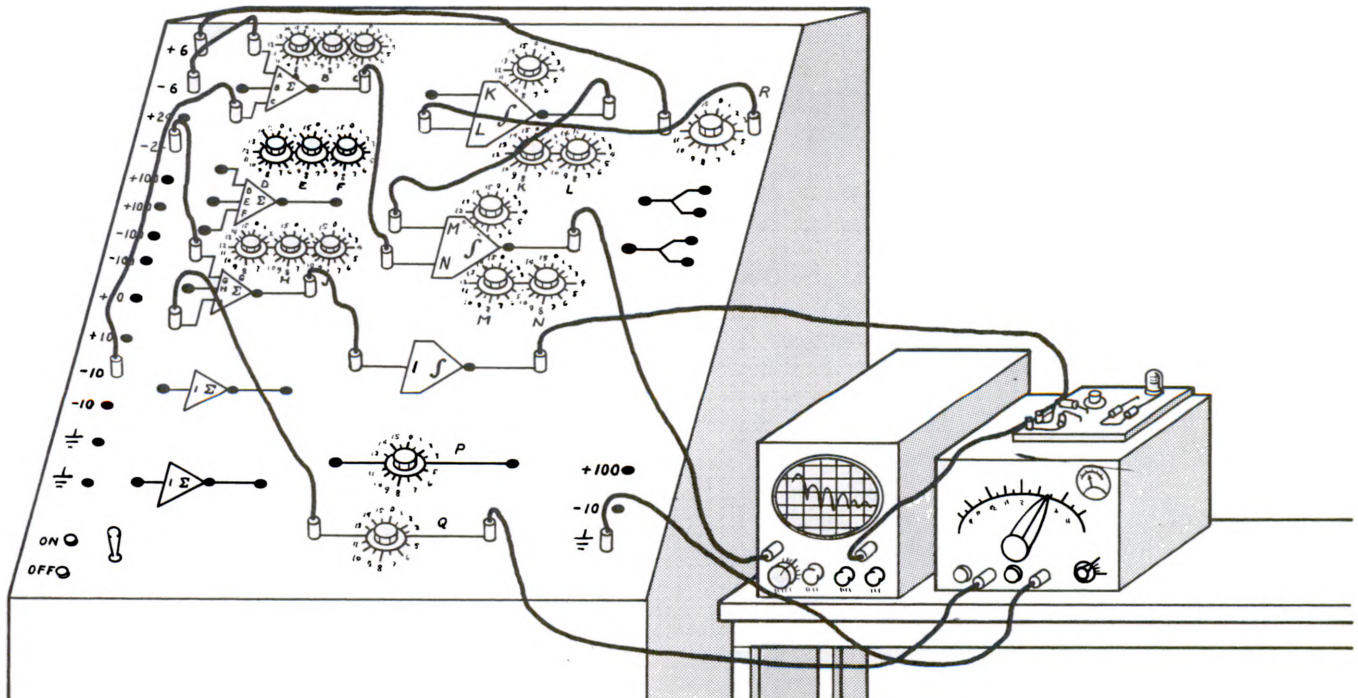
Methods of determining the existence of solutions to equations are not within the scope of this book, but the

problem illustrates the care required in programming a computer.

Once the program is completed, the machine must be set up in accordance with the road map. Most general purpose computers have the input and output terminals connected to a matrix of sockets on a panel. A patch-board, with a matrix of plugs to match the matrix of sockets, fits over the panel. The desired connections are made by wiring the points on the patchboard to correspond to the road map, and attaching the patch-board. Frequently used programs can be stored by using several patchboards.

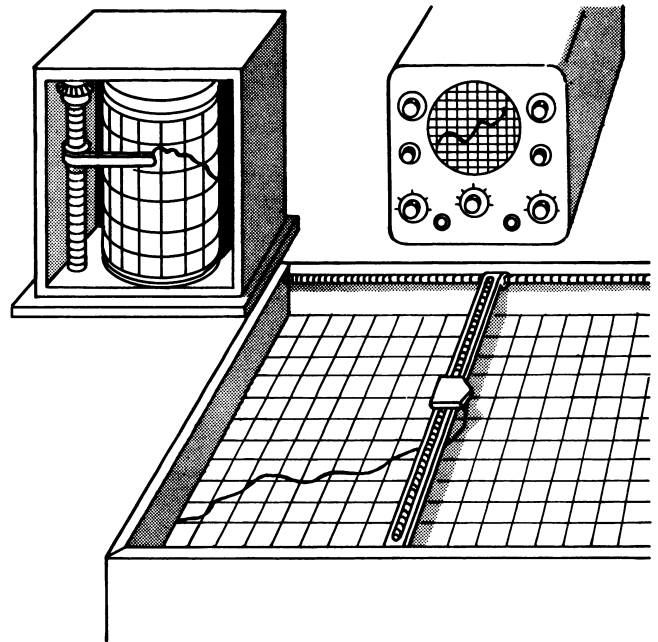
One major advantage of general purpose analog computers is that all the devices used need not be part of the computer. If some device which is not a part of the computer is to be used, those terminals to which it is to be connected are simply wired to the device instead of to other components of the computer. If a certain device is required for a problem and is not available as part of the computer, it can be obtained separately and wired into the road map. This also permits real and simulated components of a system to be used together.

Special purpose computers are not always completely pre-programmed. Some constants may be altered by resetting potentiometers or some alternate methods of computation may be provided, and connected by the operator as conditions warrant. For example, target altitude may normally be computed on a tracking computer, but if the target is known to be a ship, the operator may set the altitude at zero and ignore the computations.



form of results

Once the computer is designed and programmed, the output data must be presented in a usable form. If the output is a voltage or shaft rotation, a voltmeter or pointer may be sufficient. If the variation of the output with time is of interest, a recording voltmeter or mechanical recorder can be used. This type of instrument usually moves a pen a distance proportional to the electrical or mechanical input, while moving the paper at a constant rate. If the variations in two dimensions are required, an x, y plotter may be used. This device operates in the same manner as the simple time recorder, except that both coordinates are controlled by the input variables. Usually, the paper is stationary and the pen has two degrees of freedom. If a permanent record is not required, an oscilloscope may be used. In most cases, the results from analog computers used in weapons systems will be data which must be translated into action. In this case, the output is usually connected to some form of control system. For example, if the output were a gun elevation, the signal would be used as the order to a servo mechanism which would elevate the gun.



SUMMARY

The fundamental characteristic of an analog is similarity. The analog is similar to the subject in some area; i.e., it duplicates the geometry, dynamic behavior, or some other aspect of the subject. The basic operating principle of analog devices is continuous measurement of quantities.

Those analogs which use a different analog variable for each subject variable are called passive analogs. For example, when an inductor is used as the analog of weight, the inductance may be equivalent to mass, the current equivalent to velocity, and the voltage equivalent to force.

Those analogs in which one analog quantity can represent several different subject variables are called active analogs. For example, to duplicate a mass with an active analog which operates on rotational motion, the force would be represented as a rotational input to an integrator. Multiplied by $\frac{1}{M}$ where M is the mass, the output rotation would represent the velocity.

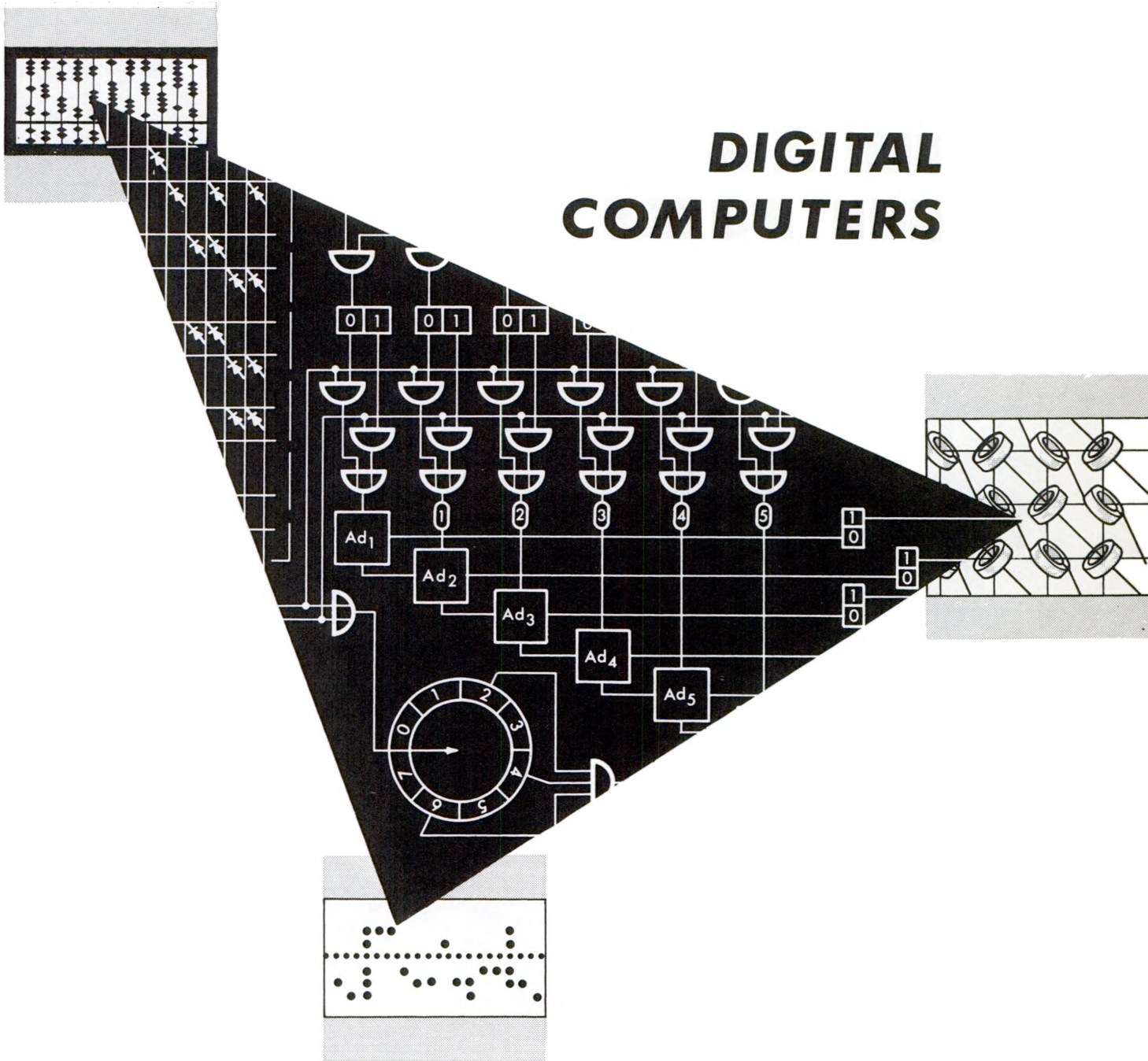
The passive analog (inductor) performs several mathematical operations simultaneously, but the various quantities are each represented by different variables.

The active analog uses a different device for each operation, but one variable (shaft rotation) may represent several different quantities. The active analog will normally require more components for a given operation than the passive analog, but is much easier to use.

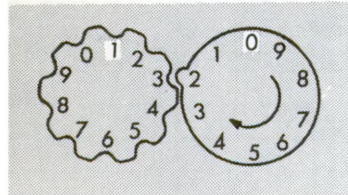
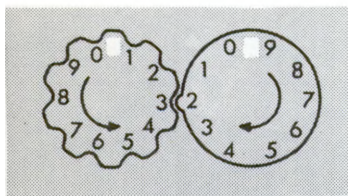
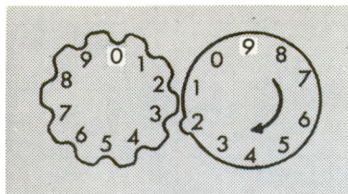
Analog computers can be employed in weapons systems wherever problems of calculations from continuous data, simulation, or control are encountered.

In the areas of target detection and tracking, analog computers are used to direct search radars and tracking devices, store data on target location and velocity, and predict future target motion. The calculations necessary to direct weapons launchers, and actually control launchers and missiles are performed by computers. In addition to actual operation of weapons, analog computers are used to simulate targets for training purposes, and evaluate the performance of weapons systems in engaging the simulated targets. The history of increasing complexity in weapons and the increasingly rapid pace of warfare, as well as improvements in the capabilities of computers, indicate a still greater dependence on computers in the future.

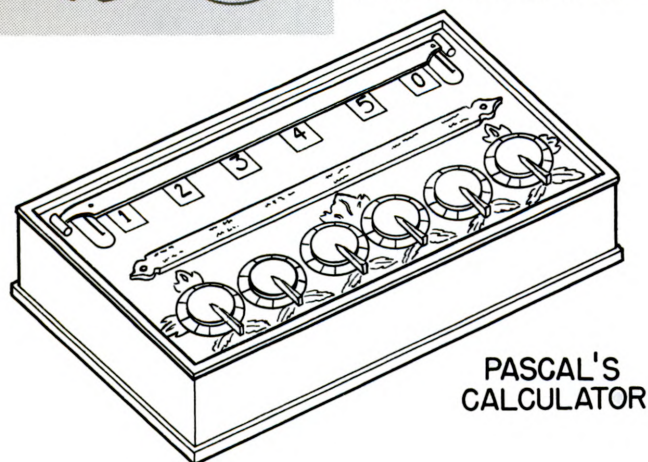
DIGITAL COMPUTERS



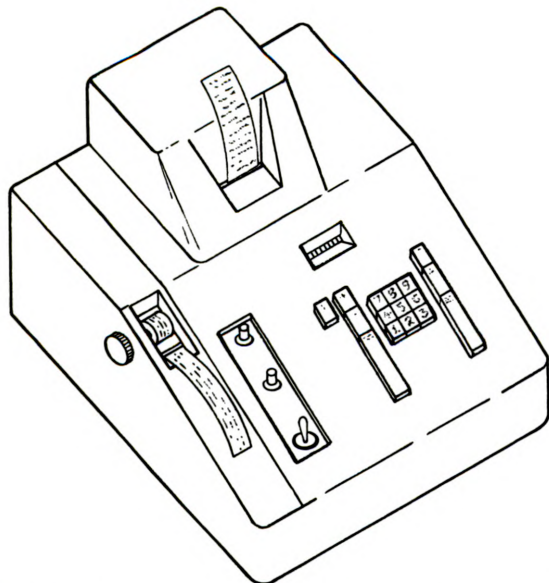
The digital computer has arisen from man's continued efforts to escape the drudgery of routine computation. Probably the earliest digital computer beyond the finger and pebble counting stage was the abacus. On an abacus, however, every operation is manual. As early as the Seventeenth Century, machines were designed by both Blaise Pascal (1642) and Baron Von Leibnitz (1672) which performed the "carry" operation automatically.



Both machines were based on ten-tooth cog wheels which caused the next more significant wheel to advance one notch when passing from nine to zero. Later machines were constructed which could multiply, divide, and subtract as well as add. On these machines the operator not only has to punch the control keys that determine which operation the machine will perform, but also has to record intermediate results. The modern desk calculator is basically the same as these earlier machines.

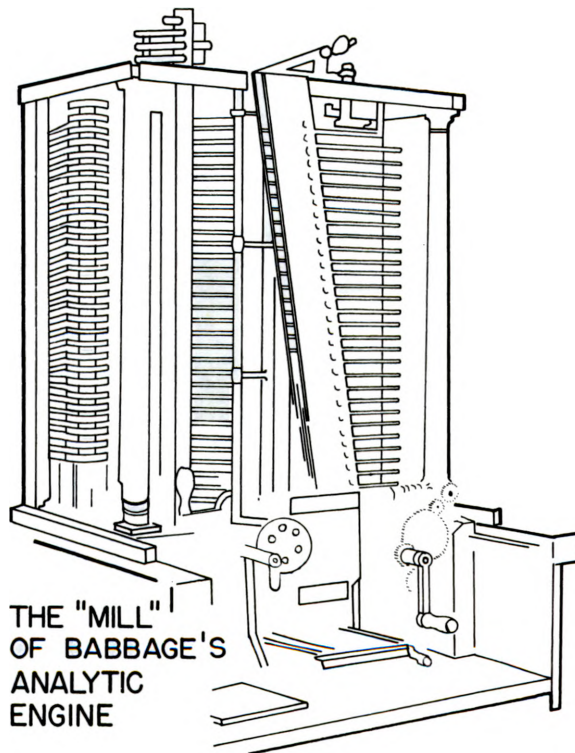


PASCAL'S
CALCULATOR



Calculators improved rapidly until 90 percent of the computation time was required to punch data and control keys. When a calculation requires many steps to reach a solution it is also necessary to record the results of each intermediate calculation. This is the major limitation of manual calculators. The operator must wait until the end of each step before he can punch the keys for the next step. He must also keep a record of the result from each step. For example, to compute

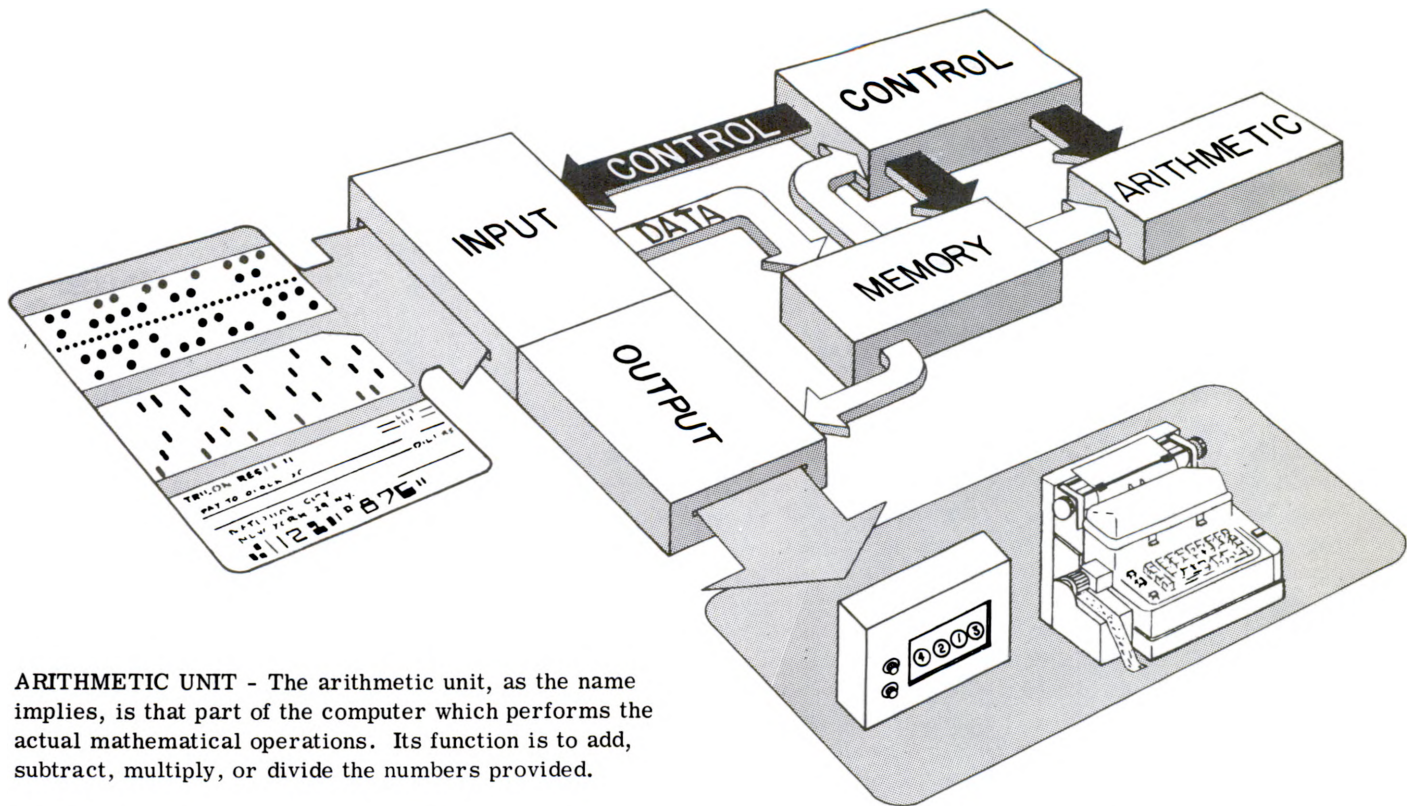
$(3+8) \div (4+2)$ on a desk calculator requires three steps. The operator must first compute $3+8$, and write down the result. This is called an intermediate result. Then he must compute $4+2$ and write down this result. Finally he must calculate the ratio of the intermediate results ($11 \div 6$). The entire operation may take a full minute, even though the calculator operates for only a second or two. The duties of the operator include providing data (the numbers involved); providing instructions (deciding what the next step will be); controlling the operations (punching the "add" or "subtract" or "multiply" or "divide" button which determines what function the machine will perform); and storing the intermediate results.



THE "MILL"
OF BABBAGE'S
ANALYTIC
ENGINE

In the 1830's, Charles Babbage attempted to construct a machine which could store data, instructions and intermediate results, and control itself. The operator would still be required to enter data and instructions, but they could all be entered at once. With this type of computer, all the data and instruction could be inserted, and the machine could run through the complete computation without depending upon an operator. Babbage's machine failed because the technology of his time was incapable of producing components with the accuracy and reliability he required. However, his basic idea is inherent in all present-day automatic digital computers. Babbage divided his machine into three parts which he called the mill, the store, and the control. A separate machine was required to communicate, i.e., provide input data and collect output data. Present automatic digital computers are divided in the same manner functionally, and sometimes physically, although the names are different. The computer consists of an arithmetic unit, and memory, control, and input-output equipment. The instructions and data are in the form of a step-by-step procedure called a program.

digital computer organization



ARITHMETIC UNIT - The arithmetic unit, as the name implies, is that part of the computer which performs the actual mathematical operations. Its function is to add, subtract, multiply, or divide the numbers provided.

MEMORY - The function of the memory is to store data, instructions, and intermediate results in a manner which permits them to be located and made available for processing.

CONTROL - The control unit is the heart of the computer. This unit interprets the instructions, sets up the arithmetic unit for the operation called for, and routes the proper data to and from the arithmetic unit and the memory.

INPUT/OUTPUT - The category of input/output equipment includes all devices necessary to communicate with the computer. The function of this equipment is to convert instructions and data into whatever form is required by the computer, and to convert the results into a usable form. In this book, little space is devoted to input/output equipment primarily because most of the devices used are fairly simple and straightforward in principle, and secondly, because there are too many different devices in common use to permit comprehensive coverage. However, it should not be assumed that the equipment is unimportant; often the largest part of a computer consists of input/output equipment. In a military computer where data is not available in advance, and is obtained from a variety of sources, the speed and accuracy requirements of this equipment are often more demanding than those of the computer proper.

The interdependence of the various parts of an automatic digital computer can be illustrated by following the solution of the same problem discussed earlier: $(3+8) \div (4+2)$. The problem would be given to the computer in the form of a step-by-step procedure called a program, which would be stored in the memory.

- | | |
|--|-------------------------|
| 1 Add the first pair of numbers | 4 Print out the answer. |
| and store the result. | 5 (3) |
| 2 Add the second pair of numbers and store the result. | 6 (8) |
| 3 Divide the result of step 1 | 7 (4) |
| by the result of step 2. | 8 (2) |

The last four parts of the program, of course, are not instructions but the data which is to be used. The control unit would interpret the first instruction, set the arithmetic unit for addition, transfer the numbers required (3 and 8) to the arithmetic unit, and transfer the result back to the memory. The arithmetic unit would perform the addition. The second step is the same as the first step, except that the data would be transferred from different memory locations, and the result stored in a different location. The third instruction would cause the control unit to set the arithmetic unit for division, and transfer the two intermediate results from the storage location used earlier. The fourth instruction would cause the control unit to transfer the answer to some output device. Note that the procedure is the same as the procedure which would be used by a man solving the same problem with a desk calculator. The only difference with the automatic computer is that the entire problem can be put into the machine at once, and the problem run through to completion without waiting for any new data or instructions. The digital computer is also like the man performing arithmetic in that it does not add real quantities according to physical laws, but rather adds symbols which represent the quantities according to fixed rules. Since digital computers deal entirely with numbers, a thorough knowledge of the structure of number systems is essential.

NUMBER SYSTEMS

It is important to distinguish between a quantity and the numeral used to represent the quantity. The quantity six may be represented in many ways (e.g., 6, ::, a half-dozen, VI, $12 \div 2$, $3 + 3$, 600×10^{-2} , $\sqrt{36}$); all are equal, but not the same. The quantities are fixed and immutable. The quantity three added to the quantity four is equal to the quantity seven, no matter what symbols are used. The numerals which represent the quantities are arbitrary. If the arithmetic process is intended to describe the process which we normally consider as addition, the result of adding that symbol which represents the quantity three to that symbol which represents the quantity four must equal that symbol which represents the quantity seven. It makes no difference what the symbols are. Throughout the remainder of this chapter, quantities are spelled out in order to distinguish them from numerals. Thus, "ten" means the quantity ten at all times, whereas "10" is a numeral which represents ten in the decimal number system but which will represent different quantities in other systems.

A decimal number, 903.5 for example, is written in "positional notation". The digits (9,0,3,5) have a value which depends on their position with respect to the decimal point. Since the decimal number system is based on powers of ten, 903.5 means $(9 \times 100) + (0 \times 10) + (3 \times 1) + (5 \times \frac{1}{10})$ or $9r^2 + 0r^1 + 3r^0 + 5r^{-1}$, where r is the radix

or base of the number system. All number systems are not based on positional notation. In the Roman number system, V means five whether it appears alone (V) or before other symbols (VIII). If the number required is fifty, another symbol must be used (L). All number systems do not use ten as the radix either. Any integer greater than 1 can be used. Some other positional notation number systems are shown below.

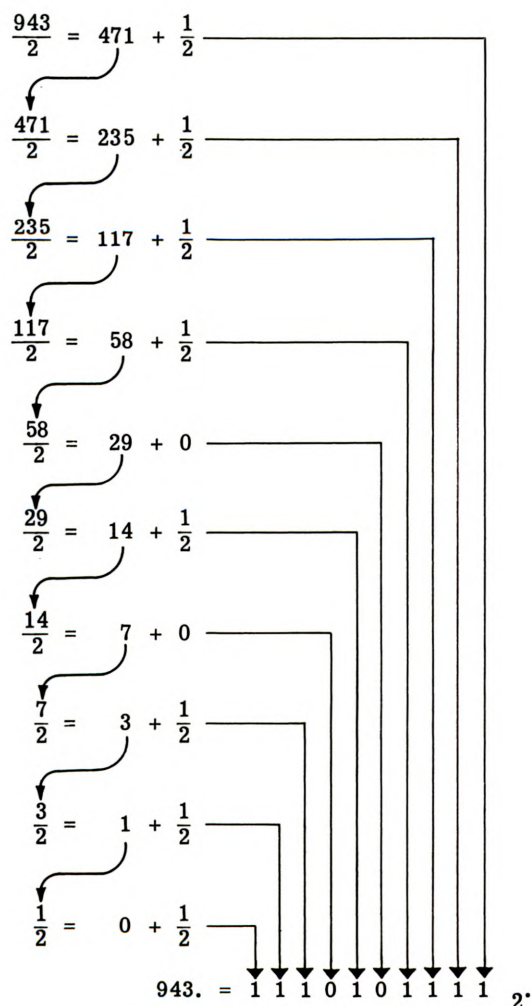
DECIMAL SYSTEM RADIX=TEN	BINARY SYSTEM RADIX=TWO	OCTAL SYSTEM RADIX=EIGHT	DUO-DECIMAL SYSTEM RADIX=TWELVE
0	0_2	0	0
1	1_2	1	1
2	10_2	2	2
3	11_2	3	3
4	100_2	4	4
5	101_2	5	5
6	110_2	6	6
7	111_2	7	7
8	1000_2	10	8
9	1001_2	11	9
10	1010_2	12	τ
11	1011_2	13	ϵ
12	1100_2	14	10
13	1101_2	15	11
14	1110_2	16	12
15	1111_2	17	13
16	10000_2	20	14
17	10001_2	21	15
18	10010_2	22	16
19	10011_2	23	17
20	10100_2	24	18
21	10101_2	25	19
22	10110_2	26	1τ
23	10111_2	27	1ϵ
24	11000_2	30	20

The subscripts are used to identify all number systems except decimal. Just as the decimal system uses ten symbols (0, 1, 2, 3, 4, 5, 6, 7, 8 and 9), the other systems use r symbols (0 through $r-1$) where r is the radix. The binary system requires only two symbols, and the duodecimal system requires twelve; so two symbols have been made up (τ = ten, ϵ = eleven). All these systems are based on the same positional notation system as the decimal system. Just as each position in a decimal number is multiplied by a power of ten ($1, 10, 100, 1000, \dots, 10^n$), each binary position is multiplied by a power of two ($1, 2, 4, 8, 16, 32, 64, \dots, 2^n$), each octal number by a power of eight ($1, 8, 64, 512, \dots, 8^n$), and each duodecimal number by a power of twelve ($1, 12, 144, 1728, \dots, 12^n$). Any number can be converted to a decimal number by performing the multiplication.

A convenient algorithm for converting from one number system to another is as follows:

For digits to the left of the radix point (radix point is a general term which includes the decimal point), divide repeatedly by the radix of the system to which the conversion is to be made, and record the remainder of each step. The remainders, in reverse order, comprise the new number.

To convert decimal 943 to binary:



To convert decimal 943 to octal:

	Remainder
$\frac{943}{8} = 117 + \frac{7}{8}$	7
$\frac{117}{8} = 14 + \frac{5}{8}$	5
$\frac{14}{8} = 1 + \frac{6}{8}$	6
$\frac{1}{8} = 0 + \frac{1}{8}$	1
$943_{10} = 1657_8$	

For digits to the right of the radix point, multiply repeatedly by the radix of the system to which the conversion is to be made, and record the digits which appear to the left of the radix point. These digits, in the order they occur, comprise the new number.

To convert decimal .640625 to octal:

	Number to Left of Radix Point
$.640625 \times 8 = 5.125000$	5
$.125 \times 8 = 1.000$	1
$.640625_{10} = .510000_8$	

To convert decimal .640625 to binary:

$.640625 \times 2 = 1.281250$	1
$.281250 \times 2 = 0.562500$	0
$.562500 \times 2 = 1.125000$	1
$.125000 \times 2 = 0.250000$	0
$.250000 \times 2 = 0.500000$	0
$.500000 \times 2 = 1.000000$	1
$.640625_{10} = .101001_2$	

This method will work for any conversion, but the division and multiplication must be performed according to the rules of the systems of the original number. Therefore, an octal number can be converted to binary by repeatedly dividing by two, but the division must be performed in the octal number system. For example, to convert 1657 to binary:

	Remainder
$\frac{1657_8}{2} = 727_8 + \frac{1}{2}$	1
$\frac{727_8}{2} = 353_8 + \frac{1}{2}$	1
etc., etc.	

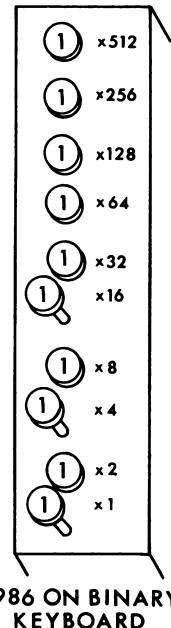
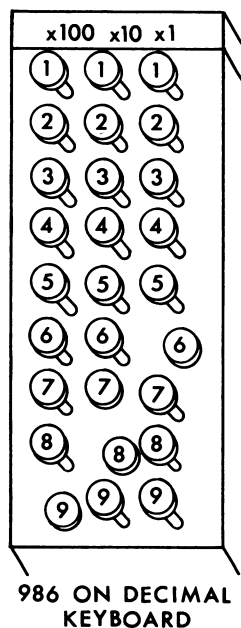
Note that the processes of arithmetic in the octal system are not the same as in the decimal system. Fortunately there is a much simpler method for octal to binary conversion. Since the largest single octal symbol (7), is equal to the largest three symbol binary number (111), simply convert each octal symbol to its binary equivalent.

OCTAL 1 6 5 7

BINARY 001 110 101 111

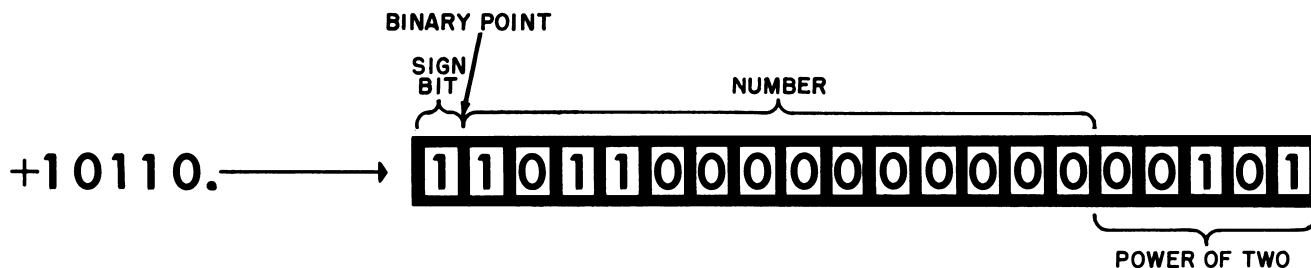
$1657_8 = 001110101111_2$

This method is valid for conversion between any number systems whose radices are integral powers or roots of one another, e.g., $8 = 2^3$. It is often easier to convert a large decimal number to the binary system by converting first to octal and then converting each octal symbol to binary. Since it is frequently used, some names have been created for the binary system. The radix point is called the binary point, and each digit is called a bit. Almost all digital computers use the binary system for internal calculations. The process of entering a number on a keyboard illustrates the economy of this system.



A 3 x 9 matrix of keys is required to record a three-digit decimal number: Since 999 is equal to 1111010111_2 in binary notation, only a 1 x 10 matrix is required for the binary keyboard, a reduction from 27 to 10 keys. The economy of elements is not peculiar to keyboards, but is of a fundamental nature. On the decimal keyboard, information is being conveyed only by the keys which are depressed. The remaining keys serve no purpose, except in the one case when zero is entered. Each key will be up nine-tenths of the time. In the case of the binary keyboard, however, each key is conveying information at all times. This principle applies to any device used to represent a number.

Just as the decimal point in a decimal number can be shifted by multiplying by a power of ten, i.e., $943 = 9.43 \times 10^2 = .943 \times 10^3$, a binary point can be shifted in the same manner by multiplying by a power of two (10_2). Therefore, $10110_2 = 101.10 \times 10_2^{10} = 10.110 \times 10_2^{11} = .10110 \times 10_2^{101}$. Since automatic digital computers represent numbers by electrical or mechanical states, it would be necessary to add some device to indicate the position of the binary point. Rather than introduce added complexity most machines treat all numbers with the binary point in the same position, usually at the left. Another convention normally adopted to reduce complexity is to require that all numbers be of the same length and in exactly the same form.



A typical arrangement for a twenty-bit capacity machine would be that the first bit represents the sign (0=-, 1=+), the following fourteen bits represents the number with the binary point at the left, and the last five bits represent the power of two. All machines do not

include provision for handling the power of two, forcing the operator to locate the binary point, just as a slide rule leaves the placement of the decimal point to the operator. Machines which include provision for the power of two are called floating-point machines.

BINARY ARITHMETIC

					ADDITION		
Addition of binary numbers is defined by the addition table:	0	1	0	1	For example:	10101	101
	+0	+0	+1	+1		+11110	110
	0	1	1	0		110011	101
				carry 1			+101
							10101

Subtraction could be performed by a similar process, but it is seldom done this way. The most frequently used method of subtraction in digital computers is the radix complement method. The radix complement of an n digit number (A) is $r^n - A$, where r is the radix of the number system.

For example, the ten's complement of 78 is: $10^2 - 78 = 22$. The nine's complement of a decimal digit (A) is equal to $9 - A$, and the nine's complement of a complete number is obtained by taking the nine's complement of each digit. The ten's complement is equal to the nine's complement, plus 1.

DECIMAL NUMBER	0 1 2 3 4 5 6 7 8 9 10 11 12
NINE'S COMPLEMENT	9 8 7 6 5 4 3 2 1 0 89 88 87
TEN'S COMPLEMENT	10 9 8 7 6 5 4 3 2 1 90 89 88
OCTAL NUMBER	0 1 2 3 4 5 6 7 10 11 12
SEVEN'S COMPLEMENT	7 6 5 4 3 2 1 0 67 66 65
EIGHT'S COMPLEMENT	10 7 6 5 4 3 2 1 70 67 66
BINARY NUMBER	0 1 10 11 100
ONE'S COMPLEMENT	1 0 01 00 011
TWO'S COMPLEMENT	10 1 10 01 100

The quickest method of obtaining the ten's complement of a number is to take the nine's complement of each digit and add 1 to the least significant digit. For example, the nine's complement of 2,496 is 7,503, and the ten's complement is 7,504; the nine's complement of 8.6 is 1.3 and the ten's complement is 1.4.

The method of subtraction for n digit numbers is derived as follows:

$$A - B = A - B \quad (1)$$

$$A - B = A - B + r^n - r^n \quad (2)$$

$$A - B = A + (r^n - B) - r^n \quad (3)$$

Since $r^n - B$ is the radix complement of B

$$A - B = A + B_{rc} - r^n \quad (4)$$

where B_{rc} indicates the radix complement of B .

The operation of subtracting r^n can be accomplished by deleting 1 from the $n+1$ st position in the decimal and binary systems. If there is no 1 in the $n+1$ st position, equation (4) can be rewritten:

$$A - B = -r^n - (A + B_{rc}) \quad (5)$$

$$A - B = - (A + B_{rc})_{rc} \quad (6)$$

Therefore, the rule for subtraction of binary or decimal numbers is to add the radix complement of the negative number to the positive number. If the result has a 1 in the $n+1$ st position, delete the 1 and the answer is positive. If the result has no 1 in the $n+1$ st position, the answer is the radix complement of the result, and is negative.

For example:

$$\begin{array}{r} 8402 \\ - 2139 \\ \hline 6263 \end{array} \quad \begin{array}{r} 8402 \\ + 7861 \\ \hline 16263 \end{array}$$

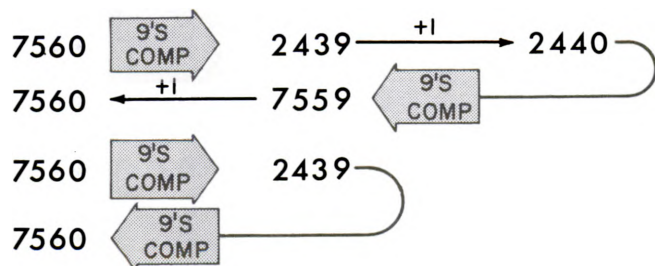
+6263 answer

$$\begin{array}{r} 2139 \\ - 8402 \\ \hline 03737 \end{array} \quad \begin{array}{r} 1597 + 1 \\ + 1598 \\ \hline 03737 \\ - 6262 + 1 \\ \hline 6263 \end{array}$$

6263 answer

The advantage of the complement method is that no circuitry for subtraction need be incorporated in the computer. The process requires only addition. It may be argued that subtraction is required to form the complement. This is true for decimal numbers, but not for binary numbers. Since the one's complement of 1 is 0, and vice-versa, it is necessary only to convert all 1's to 0's, all 0's to 1's, and add 1, in order to form the two's complement.

Since $B_{rc} = (r^n - B)$; $(B_{rc})_{rc} = r^n - (r^n - B) = B$, so complementing a number twice gives the original number.



Further simplification is obtained in most computers by recording a negative number in the two's complemented form, along with a sign bit; 0 for - and 1 for +, and with the binary point at the left of the number.

In this system, $+ .11011$ would be recorded as 1.11011 , but $- .11011$ would appear as

$$\begin{array}{r} 0.00100 \\ + .00001 \\ \hline 0.00101 \end{array}$$

When using this system, no subtraction is required. If a number is negative, it will already be recorded in the complemented form and can be simply added. If the result is negative, it will be the two's complement of the answer which is the desired form. Special provision must be made to account for the sign bit.

All six possibilities are summarized below:

SUMMARY				
	SIGN BITS	CARRY	SIGN OF ANSWER	ADD TO POWER OF TWO
1.	1 + 1	1	1	1
2.	1 + 1	0	1	0
3.	0 + 0	0	0	1
4.	0 + 0	1	0	0
5.	1 + 0	1	1	0
6.	1 + 0	0	0	0

2 and 5 may be combined, and 4 and 6 may be combined, to give:

SUM of SIGN BITS + CARRY		SIGN OF ANSWER	ADD TO POWER OF TWO
3	=	1	1
2	=	1	0
1	=	0	0
0	=	0	1

Examples:

$$\begin{array}{l} \text{eight} = +1000. = 1.1000 \times 10^{100} \\ \text{+five} = +0101. = 1.0101 \times 10^{100} \\ \hline 1.1101 \times 10^{100} = +1101 = \text{thirteen} \end{array}$$

$$\begin{array}{l} \text{eight} = +1000. = 1.1000 \times 10^{100} \\ \text{-five} = -0101. = 0.1011 \times 10^{100} \\ \hline 1.0011 \times 10^{100} = +0011. = \text{three} \end{array}$$

$$\begin{array}{l} \text{five} = +0101. = 1.0101 \times 10^{100} \\ \text{-eight} = -1000. = 0.1000 \times 10^{100} \\ \hline 0.1101 \times 10^{100} = -0011 = \text{-three} \end{array}$$

$$\begin{array}{l} \text{-five} = -0101. = 0.1011 \times 10^{100} \\ \text{-eight} = -1000. = 0.1000 \times 10^{100} \\ \hline 0.0011 \times 10^{100} = -1101. = \text{-thirteen} \end{array}$$

It has been assumed in all cases that each number appeared in the computer with the same power of two as a multiplier. This must be taken care of by the computer beforehand by shifting the smaller number to the right and reducing its exponent until the two exponents are equal.

MULTIPLICATION - Digital computers perform multiplication by repeated addition. The sign of the result is obtained just as in ordinary multiplication. If both signs are the same, the result is positive; if they are different, the result is negative. The basic process is almost exactly similar to that normally used for pencil-and-paper calculations.

PENCIL AND PAPER MULTIPLICATION

$$\begin{array}{r}
 1.1011_2 \\
 \times 1.0101_2 \\
 \hline
 11011 \\
 00000 \\
 11011 \\
 00000 \\
 11011 \\
 \hline
 10.00110111_2
 \end{array}$$

COMPUTER MULTIPLICATION

$$\begin{array}{r}
 1.0101_2 \times 1.1011_2 \\
 \hline
 \text{ADD } 11011 \leftarrow \text{SHIFT LEFT} \\
 11011 \\
 \hline
 \text{ADD } 11011 \leftarrow \text{SHIFT LEFT} \\
 11011 \\
 \hline
 10000111 \\
 \hline
 11011 \leftarrow \text{SHIFT LEFT} \\
 10000111 \\
 \hline
 \text{ADD } 11011 \leftarrow \text{SHIFT LEFT} \\
 11011 \\
 \hline
 10.00110111_2
 \end{array}$$

It is also possible to shift right and take the multiplier bits from left to right.

$$\begin{array}{r}
 1.1011_2 \times 1.0101_2 \\
 \hline
 11011 \\
 \text{SHIFT RIGHT} \rightarrow 11011 \\
 \text{SHIFT RIGHT} \rightarrow 11011 \\
 \text{SHIFT RIGHT} \rightarrow 11011 \\
 \text{SHIFT RIGHT} \rightarrow 11011 \\
 \hline
 10.00110111_2
 \end{array}$$

DIVISION - Just as repeated addition can be used for multiplication, repeated subtraction can be used for division. The number of times the divisor can be subtracted from the dividend is equal to the quotient. For example, seven can be subtracted from 21 three times; therefore, $\frac{21}{7} = 3$.

This works when the quotient is an integer (no remainder). Fractional quotients can be handled by "restoring" division. In restoring division, the divisor is subtracted until a negative remainder is reached. The previous remainder is then "restored" by adding the divisor, and the division is shifted one place to the right. The process continues until there is no remainder, or until the desired precision is attained.

Example: $22 \div 7.00$

	Count
22.00	
- 7.00	
15.00	1
- 7.00	
8.00	1
- 7.00	
1.00	1
- 7.00	
Negative remainder ∴ restore add + 7.00	0
1.00	
shift right - 0.70	
0.30	0.1
- 0.70	
Negative remainder ∴ restore add + 0.70	0.0
0.30	
shift right - 0.07	
0.23	0.01
- 0.07	
0.16	0.01
- 0.07	
0.09	0.01
- 0.07	
0.02	0.01
- 0.07	
Negative remainder ∴ stop - 0.05	0.00
	<u>3.14</u> quotient



100

100



computer codes

BINARY CODED DECIMAL

The most frequently used code in digital computers is the binary coded decimal (BCD) system (also known as "8421" code). In this system, each decimal digit, instead of an entire number, is converted to its four-bit binary form.

Decimal	8	7	1
BCD	1000	0111	0001
	Binary eight	Binary seven	Binary one

By using the system, a decimal number of any length can be accommodated, one digit at a time, by a computer with only a four-bit capacity. It is most often used where a great deal of conversion is required from decimal to binary, and vice versa, since the decimal digits can be handled one at a time. Its disadvantages are: first, it is inefficient (twelve bits are required to encode 871, only ten are required in normal binary form); second, it is not well behaved with respect to addition (for example, eight + six would give the binary form for fourteen, which is a code with no meaning in the BCD system, and sums from ten to fifteen give no carry bit to the fifth column); and third, it is not "self-complementing" (the nine's complement of eight (1000) is one, but inverting 1000 gives 0111 (seven)). Two of these disadvantages are eliminated by the excess three's binary coded decimal system.

EXCESS THREE'S		
DECIMAL	BCD	
	0000	} unused codes
	0001	
	0010	
EXCESS THREE'S		
BINARY CODED	0	0011
DECIMAL	1	0100
	2	0101
	3	0110
	4	0111
	5	1000
	6	1001
	7	1010
	8	1011
	9	1100
	1101	} unused codes
	1110	
	1111	

In this system each decimal digit of a number is represented by the binary form of the digit plus three, e.g., the excess three's BCD form for zero is binary three; for one is binary four, etc.

Examining the table shows that the system is "self complementing", e.g., the nine's complement of zero (0011) is nine (1100); the nine's complement of one (0100) is eight (1011), etc.

Since each excess three's BCD number is the binary form of the decimal number plus three, the sum of two excess three's BCD numbers is the binary form of the sum of the decimal numbers plus six.

$$A_{x3} + B_{x3} = (A+B)_2 + \text{six}$$

The binary form of sixteen is 10000, i.e., it is the smallest five-bit binary number. Therefore, the sum of any two excess three's BCD numbers which is greater than nine will give a binary number greater than fifteen and therefore generate a carry bit to the fifth position. This system is well behaved with respect to addition. Based on these two conditions, rules can be formulated for addition of excess three's BCD numbers. The individual four-bit quantities are each added according to the rules of binary arithmetic. If there is a carry into the fifth column, add three (0011) to the result (to place the result in excess three's BCD form), and add the carry in the next position. If there is no carry into the fifth column, subtract three (0011) from the result, (since $A_{x3} + B_{x3} = (A+B)_2 + \text{six}$, three must be subtracted to return to the excess three's BCD form $\{(A+B)_2 + \text{three}\}$).

Examples:

fourteen	0 1 0 0	0 1 1 1
+ sixteen	0 1 0 0	1 0 0 1
	1 0 0 1	0 0 0 0
	-0 0 1 1	+0 0 1 1
thirty	0 1 1 0	0 0 1 1
thirteen	0 1 0 0	0 1 1 0
+fourteen	0 1 0 0	0 1 1 1
	1 0 0 0	1 1 0 1
	-0 0 1 1	-0 0 1 1
twenty-seven	0 1 0 1	1 0 1 0

CYCLICALLY

PERMUTED CODE

The cyclically permuted code, (also known as reflected binary and Gray code), is frequently used when data must be converted from analog to digital form. The outstanding feature of this code is that each number differs from both adjacent numbers in only one bit.

DECIMAL	CP
0	0000
1	0001
2	0011
3	0010
4	0110
5	0111
6	0101
7	0100
8	1100
9	1101
10	1111
11	1110
12	1010

Note that changing from any number to the next higher or next lower number requires a change in only one position.

In this system, the weight of the i th position is $\pm(2^{i+1}-1)$. Just as binary positions are weighted from right to left as 1, 2, 4, 8, 16..., CP positions are weighted 1, 3, 7, 15, 31, 63.... The most significant 1 is positive, the second is negative, the third is positive, etc. For example, $101101_{CP} = +(1 \times 63) - (1 \times 15) + (1 \times 7) - (1 \times 1) = 54$. The system is used only for whole numbers.

The following algorithms will convert binary to CP and CP to binary.

BINARY TO CP CONVERSION

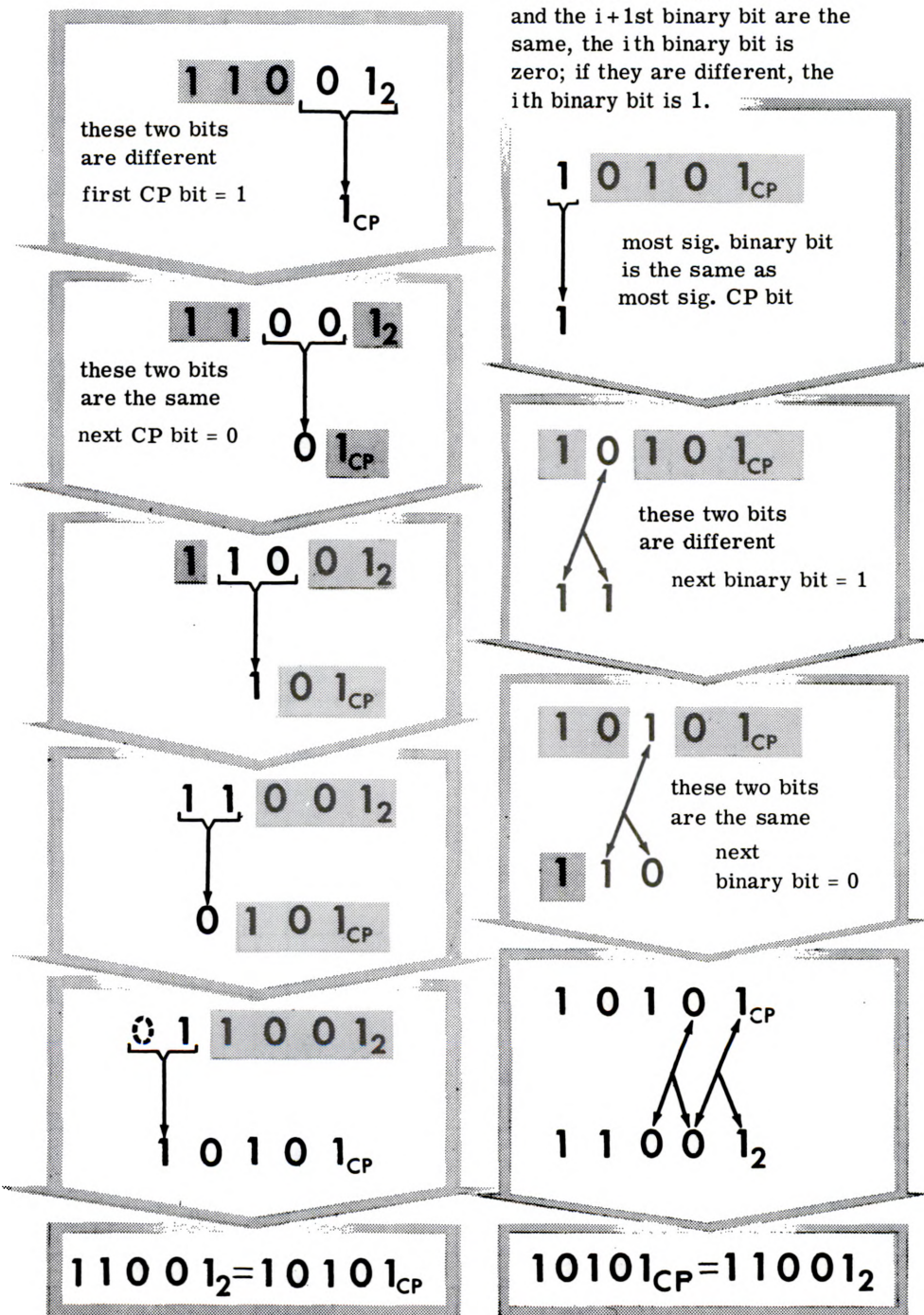
If the i th and $i+1$ st bit of the binary number are the same, the i th CP bit is 0; if they are different, the i th CP bit is 1.

CP TO BINARY CONVERSION

The most significant bit of the binary number is the same as the most significant bit of the CP number. If the i th CP bit and the $i+1$ st binary bit are the same, the i th binary bit is zero; if they are different, the i th binary bit is 1.

If conversions must be made frequently it is easier to prepare a table. Note the list of the first 12 numbers in CP code above. The right hand column, reading down, is 01100110011, etc. The next column changes from 0 to 1 after 2 rows, and thereafter every 4 rows. The third column changes from 0 to 1 after 4 rows, and thereafter every 8 rows, etc.

Many other codes are possible and some others are in frequent use. One frequently used system is the bi-quinary coded decimal. Here each decimal digit is represented by seven bits. The first two are multiples of five, the remaining five are weighted 4, 3, 2, 1, and 0, respectively. Zero is written as 1000001, i.e., $(0 \times 5) + 0$, one as 1000010, i.e., $(0 \times 5) + 1$, four as 1010000, i.e., $(0 \times 5) + 4$, five as 0100001, i.e., $(1 \times 5) + 0$, six as 0100010, i.e., $(1 \times 5) + 1$, etc. The value of this code is that each number must have exactly two 1's. By checking for this condition, any lost bits or spurious bits can be spotted immediately.



DECIMAL	BI - QUINARY WEIGHT						
	0×5	1×5	4	3	2	1	0
0	1	0	0	0	0	0	1
1	1	0	0	0	0	1	0
2	1	0	0	0	1	0	0
3	1	0	0	1	0	0	0
4	1	0	1	0	0	0	0
5	0	1	0	0	0	0	1
6	0	1	0	0	0	1	0

The binary number system and all the computer codes described have one thing in common— each bit can have only one of two values, 0 or 1. A special type of algebra, called Boolean algebra is used to handle expressions involving variables which can have only one of two values.

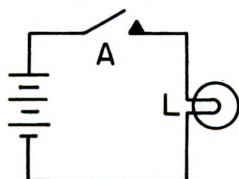
BOOLEAN ALGEBRA

Boolean algebra is a system for handling logical problems involving "true" or "false" decisions. The system handles only qualitative values such as "true", "on", or "yes", and their opposites, "false", "off",

note

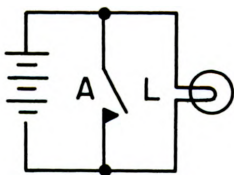
Boolean algebra is a particular topic in the "algebra of propositions" which does contain notations and rules for handling propositions such as "all A's are C's" and "some B's are C's", etc., but these are not essential to computer theory, and are not covered here.

The switch (A) and the lamp (L) in the equality circuit can be described in Boolean values. The closed position of the switch is represented by $A = 1$; the open position is represented by $A = 0$. Similarly, the on condition of the lamp is represented by $L = 1$, and the off condition by $L = 0$. The condition of the lamp is a function of the position of the switch. The equation is $L = A$. When $A = 1$ (switch closed), $L = 1$ (lamp on), and when $A = 0$ (switch open), $L = 0$ (lamp off). Note that nothing is said about how much current flows through the lamp, or about how much light is produced. Only the on or off conditions are recognized.



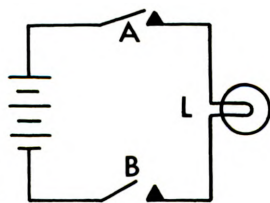
EQUALITY CIRCUIT

The not circuit is the inverse of the equality circuit; the lamp is off when the switch is closed, and vice versa. The Boolean equation for this function is $L = \bar{A}$, read as L is equal to NOT A. (The symbols \bar{A} , A , and A' are often used instead of \bar{A} .) When $A = 1$, $L = 0$, and when $A = 0$, $L = 1$.



NOT CIRCUIT

In the AND circuit, the condition of the lamp is a function of the position of both switch A and switch B. The lamp will light only when both switches are closed. The Boolean equation for this function is $L = AB$, read L is equal to A AND B. (The symbols $A \wedge B$, $A B$, $A \& B$, and $A \times B$ are also used.) $L = 1$ when $A = 1$ and $B = 1$, and only for this combination. This is summarized in the truth table for the AND function. The truth table shows the condition of the dependent variable (L) for each possible combination of the independent variables (A, B).



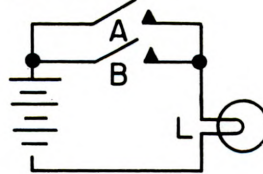
AND CIRCUIT

A	B	$L = AB$
0	0	0
0	1	0
1	0	0
1	1	1

TRUTH TABLE

or "no". Quantitative values expressing "how much" or "how many" are not involved. Conventionally, propositions which are "true", "on", or "yes" are represented as 1, and propositions which are "false", "off", "no", etc., as 0.

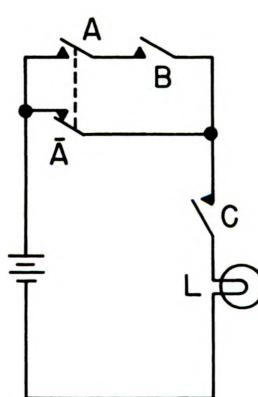
The fourth basic function is the OR function. In this case the lamp will light if either switch A or switch B (or both) are closed. The Boolean equation is $L = A + B$, read L is equal to A OR B. (The symbol $A \vee B$ is also used.)



OR CIRCUIT

A	B	$L = A + B$
0	0	0
0	1	1
1	0	1
1	1	1

More complex functions can be described in the same manner.



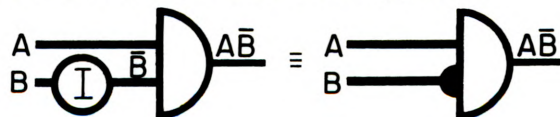
$$L = C(AB + \bar{A})$$

A	B	C	L
0	0	0	0
0	0	1	1
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	0
1	1	0	0
1	1	1	1

In addition to the equations, a system of diagrams is used. The three basic symbols represent the devices used to implement a logical proposition in a computer, and are called an AND gate, OR gate and inverter.



The presence of an inverter is often indicated by a dot or circle at the input to the following gate.



The rules for Boolean "addition" and "multiplication" can be formed from the truth tables for the AND function and OR function.

$AB = 1$ only when both $A = 1$ and $B = 1$.

Therefore:

$$\begin{array}{ll} 0 \times 0 = 0 & 0 \times 1 = 0 \\ 1 \times 0 = 0 & 1 \times 1 = 1 \end{array}$$

$A + B = 1$, except when both $A = 0$ and $B = 0$.

Therefore:

$$\begin{array}{ll} 0 + 0 = 0 & 0 + 1 = 1 \\ 1 + 0 = 1 & 1 + 1 = 1 \end{array}$$

$\bar{A} = 0$ when $A = 1$, and vice versa.

Therefore:

$$\bar{0} = 1 \quad \bar{1} = 0$$

Consider the proposition $A + \bar{A}$. It is evident that either A or \bar{A} must be 1. Therefore, this proposition is always 1. Similarly, the proposition $A \bar{A}$ must always be zero. This gives the first two postulates for manipulating Boolean equations:

1. $A + \bar{A} = 1$
2. $A \bar{A} = 0$

In considering "true" or "false" propositions it is obviously immaterial in which order these operations are performed. Therefore, the commutative and associative rules of addition and multiplication must hold and yield four more postulates.

3. $A + B = B + A$
4. $AB = BA$
5. $A + (B + C) = (A + B) + C = A + B + C$
6. $A(BC) = (AB)C = ABC$

The following nine postulates can be verified by preparing truth tables based on the rules of "addition" and "multiplication".

7. $A + 0 = A$
8. $A + 1 = 1$
9. $A + A = A$
10. $A0 = 0$
11. $A1 = A$
12. $AA = A$
13. $AB + AC = A(B + C)$
14. $A + BC = (A + B)(A + C)$
15. $\bar{\bar{A}} = A$

These postulates may be used to derive another useful relationship.

$$\begin{array}{ll} 16. & A(A+B) = A \\ & A(A+B) = AA + AB \quad \text{rule 14} \\ & = A + AB \quad \text{rule 12} \\ & = A(1+B) \quad \text{rules 11 and 13} \\ & = A1 \quad \text{rule 8} \\ & = A \quad \text{rule 11} \end{array}$$

The final (and most useful) two postulates are known as De Morgan's Theorem. Any Boolean expression is equal to the NOT function of the expression obtained by changing all AND functions to OR functions, and vice versa, and replacing each variable with its NOT function.

17. $A + B + C = \overline{(\bar{A}\bar{B}\bar{C})}$
18. $ABC = \overline{(\bar{A} + \bar{B} + \bar{C})}$

The usefulness of De Morgan's theorem is demonstrated by the proof of the equation:

$$\overline{(AB+BC+CA)} = \bar{A}\bar{B} + \bar{B}\bar{C} + \bar{C}\bar{A}$$

$$\begin{array}{ll} \overline{(AB+BC+CA)} = (\bar{A} + \bar{B})(\bar{B} + \bar{C})(\bar{C} + \bar{A}) & \text{rule 17} \\ (\bar{A}\bar{B} + \bar{A}\bar{C} + \bar{B}\bar{B} + \bar{B}\bar{C})(\bar{C} + \bar{A}) & \text{rule 14} \\ (\bar{A}\bar{B} + \bar{A}\bar{C} + \bar{B})(\bar{C} + \bar{A}) & \text{rules 12 and 8} \\ \bar{A}\bar{B}\bar{C} + \bar{A}\bar{B}\bar{A} + \bar{A}\bar{C}\bar{C} + \bar{A}\bar{C}\bar{A} + \bar{B}\bar{C} + \bar{B}\bar{A} & \text{rule 14} \\ \bar{A}\bar{B}\bar{C} + \bar{A}\bar{B} + \bar{A}\bar{C} + \bar{A}\bar{C} + \bar{B}\bar{C} + \bar{B}\bar{A} & \text{rule 12} \\ \bar{A}\bar{B}\bar{C} + \bar{A}\bar{B} + \bar{A}\bar{C} + \bar{B}\bar{C} & \text{rule 9} \\ \bar{A}\bar{B}(\bar{C} + 1) + \bar{A}\bar{C} + \bar{B}\bar{C} & \text{rule 13} \\ \bar{A}\bar{B} + \bar{A}\bar{C} + \bar{B}\bar{C} & \text{rule 8} \end{array}$$

The use of this type of notation is essential, since the language cannot handle complex logical statements without considerable ambiguity. We can say A or B and C , but it is not clear whether the meaning is A OR (B AND C), or (A OR B) AND C .

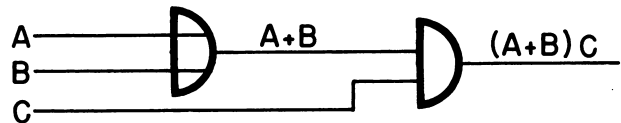
COMPUTER LOGIC DIAGRAMS

$$\rightarrow \text{AND}, \rightarrow \text{OR}, \text{---} \text{NOT}, A + (BC)$$

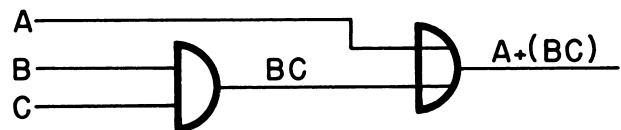
The use of the symbols

is essential also, since a large list of equations is difficult to follow.

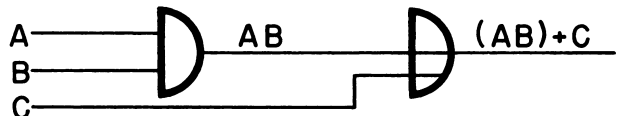
The statement $(A + B)C$ would be diagrammed as:



the statement $A + (B C)$ as:



and the statement $(A B) + C$ as:



The principal use of Boolean algebra is in the design of devices to implement the various logical requirements of a computer. For example, consider the design of a device to add two binary bits, A and B . The truth table:

A	B	SUM	CARRY
0	0	0	0
0	1	1	0
1	1	0	1
1	0	1	0

shows the value of the sum and carry for all possible combinations of A and B .

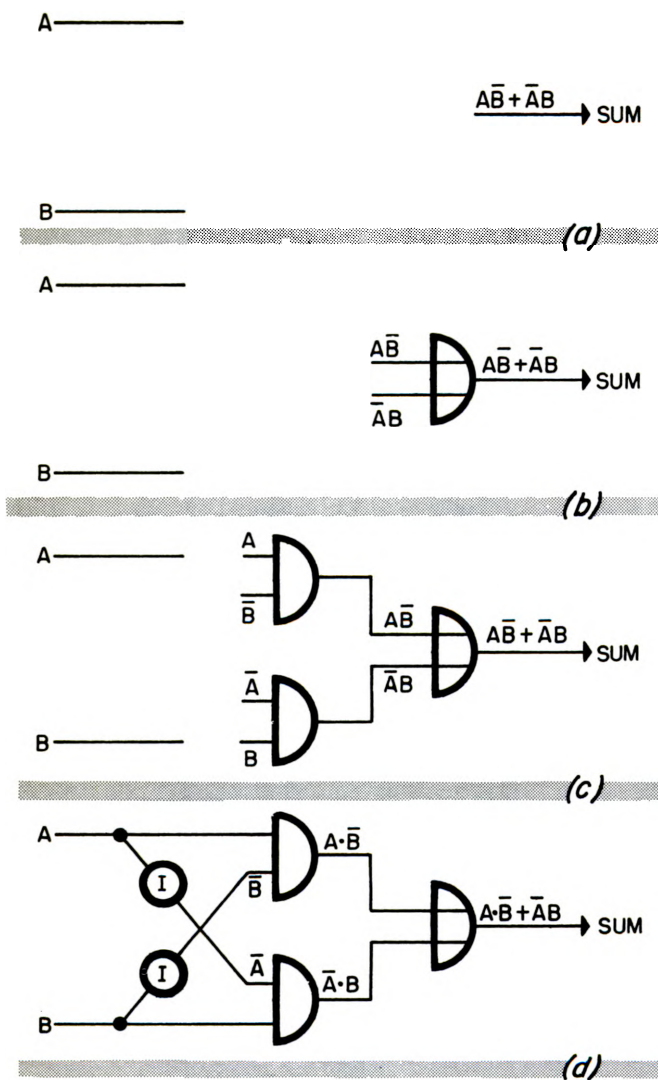
The truth table leads directly to the equations. The third line shows that the carry is 1 when $A = 1$ AND $B = 1$. Therefore:

$$\text{CARRY} = AB$$

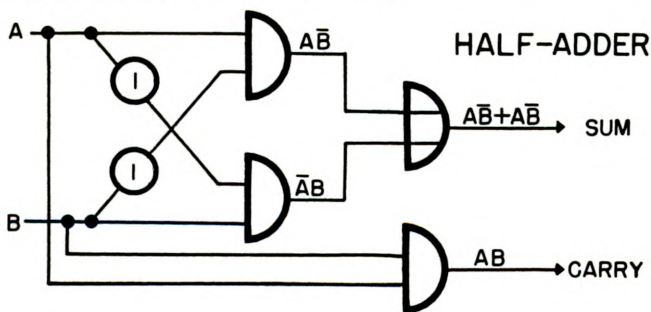
Similarly, the second and fourth lines show that the sum is 1 when $(A = 1 \text{ AND } B = 0)$ OR $(A = 0 \text{ AND } B = 1)$. Therefore:

$$\text{SUM} = A\bar{B} + \bar{A}B$$

Going from the equation for the sum to the logic diagram:



using the equation for the carry,



gives the logic diagram for the half-adder. A full-adder is a device which can add two binary bits plus a carry from the previous position, and can be made from two half-adders.

The equations for a full-adder:

$$\begin{aligned} \text{SUM} &= A\bar{B}\bar{C} + \bar{A}B\bar{C} + \bar{A}\bar{B}C + ABC \\ \text{CARRY OUT} &= AB\bar{C} + \bar{A}BC + A\bar{B}C + ABC \quad (C = \text{CARRY IN}) \end{aligned}$$

can be manipulated to be compatible with the half-adder outputs.

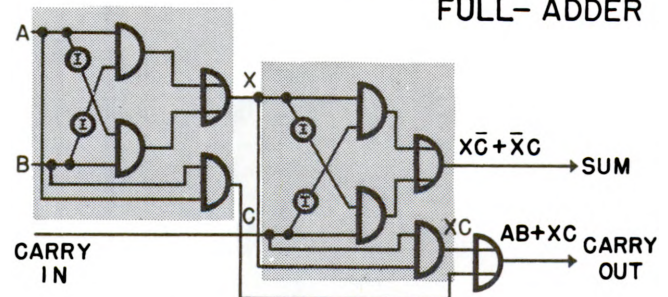
$$\begin{aligned} \text{SUM} &= (\bar{A}B + A\bar{B}) \bar{C} + (\bar{A}\bar{B} + AB) C \\ &= (\bar{A}B + A\bar{B}) \bar{C} + (A + B) (\bar{A} + \bar{B}) C \quad \text{rules 17 and 18} \\ &= (\bar{A}B + A\bar{B}) \bar{C} + (\bar{A}\bar{A} + A\bar{B} + \bar{A}B + B\bar{B}) C \quad \text{rule 14} \\ &= (\bar{A}B + A\bar{B}) \bar{C} + (0 + A\bar{B} + \bar{A}B + 0) C \quad \text{rule 2} \\ &= (\bar{A}B + A\bar{B}) \bar{C} + (A\bar{B} + \bar{A}B) C \\ &= X\bar{C} + \bar{X}C \end{aligned}$$

where $X = \bar{A}B + A\bar{B}$ and is the sum output of a half-adder. Compare this with the "sum" equation for the half-adder.

$$\begin{aligned} \text{CARRY OUT} &= AB\bar{C} + \bar{A}BC + \bar{A}BC + ABC \\ &= AB(\bar{C} + C) + C(\bar{A}B + AB) \quad \text{rule 13} \\ &= AB + C(\bar{A}B + AB) \quad \text{rule 1} \\ &= AB + CX \end{aligned}$$

The full-adder follows from these two equations.

FULL- ADDER



The basic logical connectives (AND and OR) are considered fundamental simply because they are expressed in a single word. Other combinations are also possible. For example: neither A nor B ($\bar{A} \bar{B}$); or not both A and B ($\bar{A} + \bar{B}$). Logic using these connectives is called nor logic.

NOR LOGIC

The truth table for the AND function, OR function and the two basic functions used in nor logic is shown below:

A	B	A AND B	A OR B	NEITHER A NOR B	NOT BOTH A AND B
0	0	0	0	1	1
0	1	0	1	0	1
1	0	0	1	0	1
1	1	1	1	0	0

Note that the function "neither A nor B" is the inverse of the OR function. This is called the NOR function, and is written as $A \downarrow B$. (Read as A NOR B, or as A Pierce B. This is frequently called Pierce's function, named after the man who first proposed its use.)

$$A \downarrow B = \overline{A + B} = \bar{A} \bar{B}$$

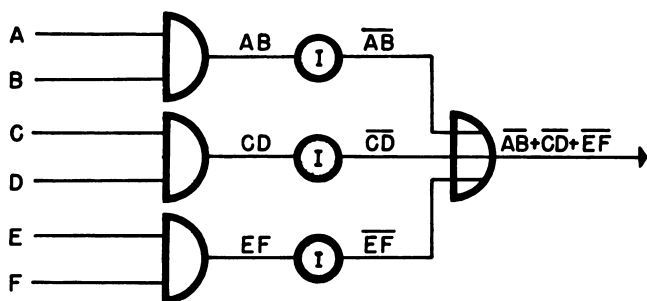
The expression "not both A and B" is the inverse of the AND function. It is called the NAND function, and is written as A / B . (Read as A NAND B, or as A stroke B. This is sometimes called the "Sheffer stroke" function, also named after the originator.)

$$A / B = \overline{AB} = \bar{A} + \bar{B}$$

The symbol is easily remembered. The AND function is written in the same manner as multiplication (AB); and its inverse, the NAND function, is written in the same manner as division (A/B) which is the opposite of multiplication.

This type of logic is frequently used because it is sometimes simpler to mechanize. For example, suppose a circuit is available which will produce a positive output (C) whenever both inputs (A, B) are zero. If a positive voltage is considered as a logical 1, the Boolean equation is $C = \bar{A} \bar{B} = \bar{A} + \bar{B}$. This is a NOR gate. By adding

A logical equation in the form $\bar{A}\bar{B} + \bar{C}\bar{D} + \bar{E}\bar{F}$ is said to be in AND/OR form. This type of equation can be converted to NAND/NOR form by using De Morgan's theorem. The logic diagram for this equation uses the same number of gates and fewer inverters than same equation in AND/OR form. For a given type of circuitry, one particular form will usually lead to the most economical design.



an inverter it can be made into an OR gate, but additional components will be required. Since it is just as easy to work with NOR gates, the added complexity is often wasteful. This particular situation arises more often than not with some types of electronic devices which are described later.

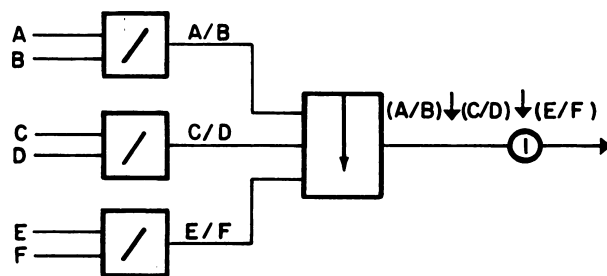
$$\bar{A} \bar{B} + \bar{C} \bar{D} + \bar{E} \bar{F} = \overline{(\bar{A} \bar{B}) (\bar{C} \bar{D}) (\bar{E} \bar{F})}$$

$$\text{Since } (\bar{X} + \bar{Y}) = X/Y$$

$$\bar{A} \bar{B} + \bar{C} \bar{D} + \bar{E} \bar{F} = (\bar{A}/\bar{B}) (\bar{C}/\bar{D}) (\bar{E}/\bar{F})$$

$$\text{Since } (\bar{X}) (\bar{Y}) (\bar{Z}) = X + Y + Z$$

$$\bar{A} \bar{B} + \bar{C} \bar{D} + \bar{E} \bar{F} = (\bar{A}/\bar{B}) \downarrow (\bar{C}/\bar{D}) \downarrow (\bar{E}/\bar{F})$$



These four functions (AND, OR, NAND, NOR) are not the only possible functions. There are, in fact, sixteen possible functions of two variables, all of which are shown here.

A	B	f ₁	f ₂	f ₃	f ₄	f ₅	f ₆	f ₇	f ₈	f ₉	f ₁₀	f ₁₁	f ₁₂	f ₁₃	f ₁₄	f ₁₅	f ₁₆
0	0	0	1	0	0	0	1	1	1	0	0	0	1	1	1	0	1
0	1	0	0	1	0	0	1	0	0	1	0	1	1	1	0	1	1
1	0	0	0	0	1	0	0	1	0	0	1	1	1	0	1	1	1
1	1	0	0	0	0	1	0	0	1	1	1	0	0	1	1	1	1

NOR ↑ AND EQUALITY ↑ NAND EXCLUSIVE OR ↑

The first is always 0 and last is always 1. These convey no information so they are of no use. The eighth combination is the equality function: It is 1 when A and B are the same (both 0 or both 1). The eleventh is the inverse of the equality function. These two can be used, but are seldom of practical value. The remaining connectives other than the four in common use are not commutative. Note that the value for (A = 1, B = 0) is different than the value for (A = 0, B = 1) in all the remaining combinations. The value of the function depends upon the order in which the variables appear. These are sometimes used for computers designed specifically to handle problems involving branches of mathematics dealing with variables having this property. Ordinary arithmetic is strictly commutative; however, (A plus B = B plus A, A times B = B times A, etc.). The connectives normally used are AND, OR, and their inverse functions NAND and NOR. De Morgan's theorem also applies to the NAND and NOR functions. Any expression is equal to the NOT function of the expression obtained by changing each NAND to NOR, each NOR to NAND, and each variable to its NOT function.

For example:

$$(\bar{A} \downarrow B) / (C \downarrow \bar{D}) = \overline{(\bar{A}/\bar{B}) \downarrow (\bar{C}/\bar{D})}$$

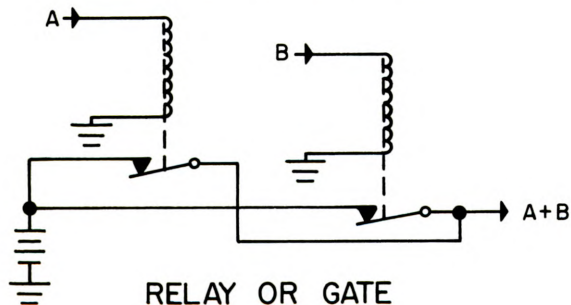
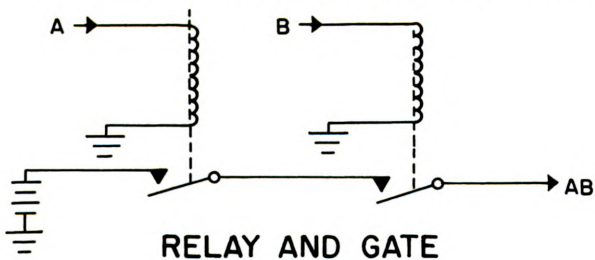
Unlike the analog computers, wherein the components may be electronic, mechanical, hydraulic or one of many other forms, automatic digital computer logic is almost exclusively electronic. This is due primarily to the fact that a given operation usually requires many more steps in a digital computer than in an analog computer. Addition on an analog computer, for example, can be performed almost instantly by a differential gear. The digital adder requires that a signal pass through two gates and an inverter to add a single pair of bits. Ten bits are required for three-decimal precision, so one addition may require thirty operations, not counting steps necessary to transfer data to and from the arithmetic unit, nor steps required to interpret the instructions. Furthermore, addition is the simplest operation in most digital computers. Obviously, if the solution time is to be kept within reason, each step must be performed very rapidly. Even the slowest present-day automatic digital computers perform 1,000 to 50,000 steps per second. Such speeds are rarely attainable with any but electronic components. That non-electronic digital devices are available is evident from the many types of mechanical desk calculators, but these are not automatic.

ELECTRONIC DIGITAL DEVICES

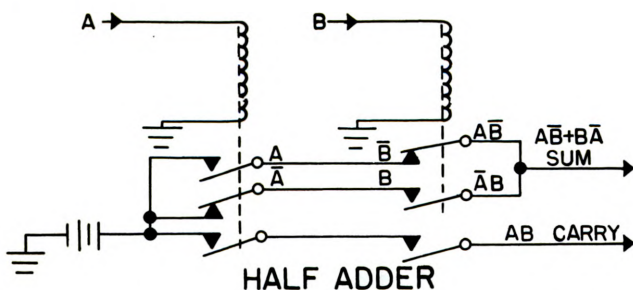
There are many electronic devices with the two-state nature necessary for binary logic. Relays may be open or closed, diodes pass current in one direction but not in the other, vacuum tubes and transistors may be conducting or non-conducting, magnetic cores may be magnetized clockwise or counterclockwise. These devices account for the vast majority of digital computer circuits. In addition, there are a variety of exotic devices with unique characteristics. The cryotron, magnetic film, tunnel diode, twistor, and aperturated plate have been applied to digital circuits.

relays

The simplest and cheapest of these devices is the relay. A relay AND gate consists of two relays, both of which must be closed in order for an output to be present. The relay OR gate uses two relays in parallel. One of the major advantages of relays is that a single relay can control many contacts (frequently as many as 100).



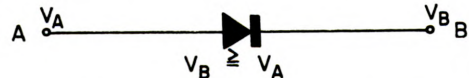
The schematic for a relay half-adder is shown. Note that the use of multiple contacts allows a logic diagram consisting of four gates and two inverters to be mechanized by only two relays. Following the logic through the schematic will make the value of a logic diagram evident.



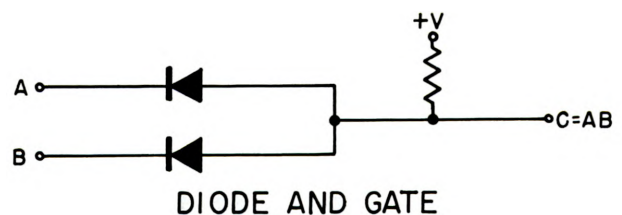
Latching relays are available which remain closed until deliberately reset even if the actuating current is removed, permitting the use of relays as a memory device. Relays are inexpensive, easily maintained, pass signals with very little loss, and are relatively insensitive to adverse humidity. However, they are usually too bulky for use in complex computers, and have a short life in terms of number of on-off cycles.

diodes

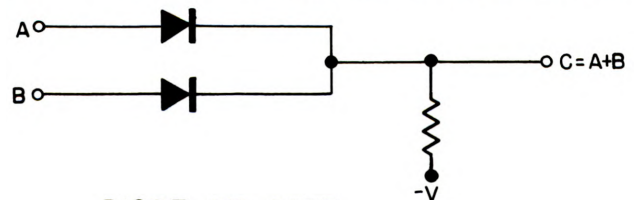
The diode is another device frequently used in digital logic. It will offer no resistance to current in the



direction of the arrow on the symbol, and nearly infinite resistance to current flow opposed to the direction of the arrow. The voltage at point B can be higher than at point A, but cannot be lower, since sufficient current would flow from A to B to eliminate any positive voltage difference. With two diodes arranged as shown, the voltage at C will be high if both A and B are high. If the voltage at either A or B is zero, the voltage at C will be zero. If a positive voltage is considered a logical 1, this is an AND gate.



If the diodes are reversed, the output voltage will be high if either A or B is high. This is a logical OR gate.

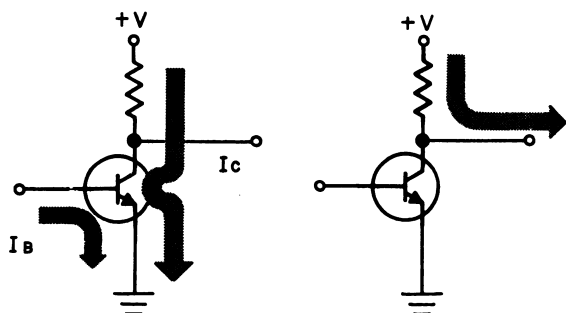


It is natural to think of a positive voltage as a logical 1, and zero voltage as a logical 0. It is not necessary, and frequently not desirable, to define it this way. It is equally valid to define a positive voltage as a logical zero, and zero voltage as a logical 1, or any other combination of positive, negative or zero voltage. If the higher voltage is considered a logical 0, the AND gate just described is an OR gate, and vice versa.

The semiconductor diode is extremely compact. If voltages are kept small, it is relatively long-lived. Its major disadvantage is that the resistance to forward current is never quite zero, nor the resistance to back current quite infinite. Thus, a signal grows slightly weaker each time it passes a gate. Vacuum tube diodes reduce this problem, but at the cost of greatly increased cost and size.

transistors

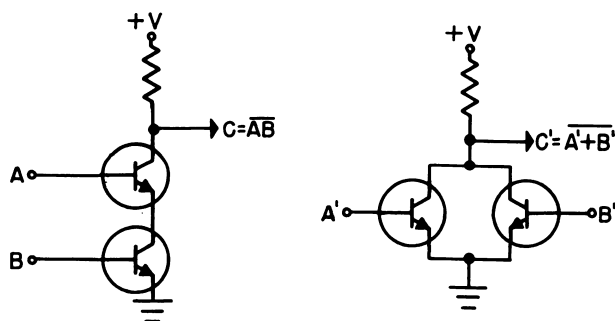
The transistor is another very important device in electronic digital computers. The important characteristic of the transistor in digital applications is that current flow into the base (in the direction of the arrow on the transistor symbol) will cause the transistor to conduct current from the collector to the emitter. (The transistor actually conducts electrons the other way, of course; hence the names emitter and "collector".) This is an "NPN" transistor. The "PNP" transistor (with the arrow in the opposite direction) operates in the same manner when negative current is applied at the base.



When the transistor is not conducting (input current is zero), current from the supply (+V) flows in the output line. When an input current causes the transistor to conduct, all current will flow through the transistor and none will flow in the output line. This is a logical inverter.

A logic gate can be constructed using two transistors connected as illustrated. Current will flow in the output line (C), if either transistor is non-conducting, i.e., if either A or B is zero. The output current will be zero if both transistors are conducting, i.e., if there is current at both A and B.

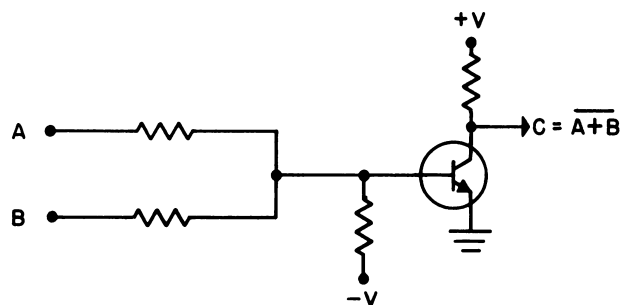
If positive current is considered a logical 1, the Boolean equation is $C = \bar{A} + \bar{B}$ or $C = \overline{AB}$. This is the AB equation for a NAND gate. By adding an inverter, the equation becomes $D = AB$. This is a logical AND gate, but requires three transistors as opposed to two for the NAND gate.



Connecting the transistors differently, there will be current at C' only if both transistors are not conducting ($A' = 0$ and $B' = 0$). The Boolean equation is $C' = \bar{A'} \bar{B'} = A + B$. This is a logical NOR gate. By

adding an inverter, this becomes an OR gate. This reveals the purpose of introducing nor logic. The use of conventional logic in transistor circuits of this type will require approximately 50 percent more transistors than nor logic. Vacuum tube gates operate in a similar manner, with nor logic having the same advantage.

This type of logic is called direct-coupled transistor logic. Each transistor is directly connected to the following stage. A transistor with a pair of resistors connected to the base will also act as a gate.

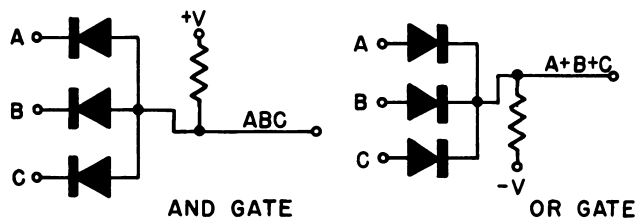


The resistors may be chosen so that current through any one resistor is sufficient to cause the transistor to conduct. Since the output is zero when the transistor is conducting, this is a NOR gate. This is called resistor-coupled transistor logic, or transistor resistor logic (TRL). Note that a basic function of an OR gate is not to provide an output when either input is present. This could be accomplished by simply connecting the input lines together. The gate is needed to prevent signals on one line from being transferred to other lines. For example, if two lines (A,B) were simply tied together, the output (C) would be 1 if either A or B or both were 1. The OR function would be correctly executed. However, if A were 1 and B were 0, the current from A would flow not only through C, but also back through line B. Thus, all other gates which used B as an input would receive current, although B = 0. Therefore, the OR gate is required to prevent this erroneous coupling of signals.

An analogous situation arises if an attempt is made to construct a TRL NAND gate. The NOR gate just described will not introduce errors of this type, because the presence of a signal in one line causes the transistor to conduct, and all current will pass through the transistor; no current (or very little) will pass through the other input resistors. A NAND gate could be constructed by making the input resistor large enough so that current in all inputs is required to make the transistor conduct. However, if current were present in only one input line, the transistor would not conduct, and hence would present a high resistance. Current flowing in one resistor would then flow out the other resistor, creating an erroneous signal in the second line. For this reason, logic circuits of this type employ only NOR gates and inverters. All NAND functions can be mechanized by NOR gates and inverters, according to De Morgan's theorem.

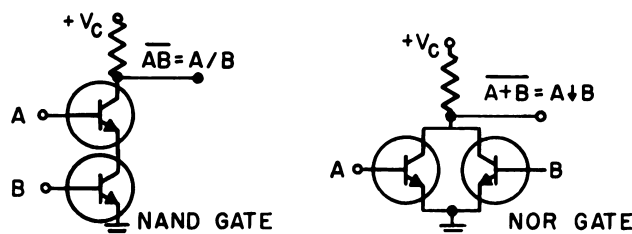
Other types of transistor logic are also used. The basic gates used with each type of diode and transistor logic are illustrated. Those which give NOR gates can be used with conventional AND/OR logic by adding another transistor inverter. Each type has some advantages, and incurs some disadvantages. Vacuum tube circuits operating in a similar manner are also possible. They have the advantage of lower initial cost, and permit the use of higher voltages. However, for large-scale computers, vacuum tube circuits are usually too bulky or unreliable.

DIODE LOGIC (DL)



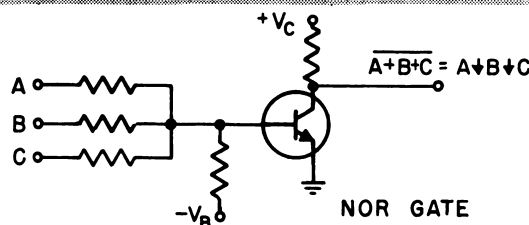
Operation is described in the text. Basic functions are AND and OR. Features high speed with inexpensive components. Since diodes are not perfect, having some finite forward resistance, signals are attenuated at each gate, making periodic amplification necessary. Noise rejection is poor.

DIRECT COUPLED TRANSISTOR LOGIC (DCTL)



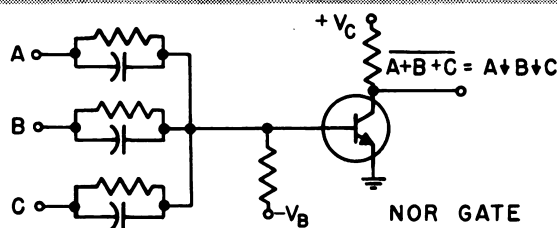
Operation is described in the text. Circuits are designed so the output signal is only slightly dependent on the size of the input signal. Therefore, substandard pulses will be restored by the gate, and no amplification is necessary. High speed, low voltage operation is practical.

RESISTOR TRANSISTOR LOGIC (RTL)



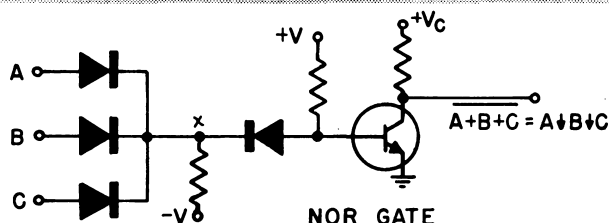
Operation is described in the text. When the transistor is conducting, it will develop a surplus of carriers in the base ("hole storage"). When the input signal is removed, the transistor will continue to conduct until this charge is dissipated, thereby limiting the speed of operation.

RESISTOR CAPACITOR TRANSISTOR LOGIC (RCTL)



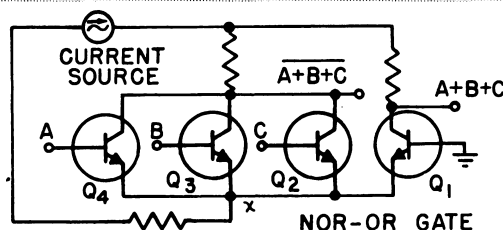
Operation is basically the same as RTL except that the capacitors increase the initial base current, reducing the turn-on time. They also build up a charge which neutralizes the charge stored in the base when the signal is removed, reducing the cut-off time.

LOW LEVEL LOGIC



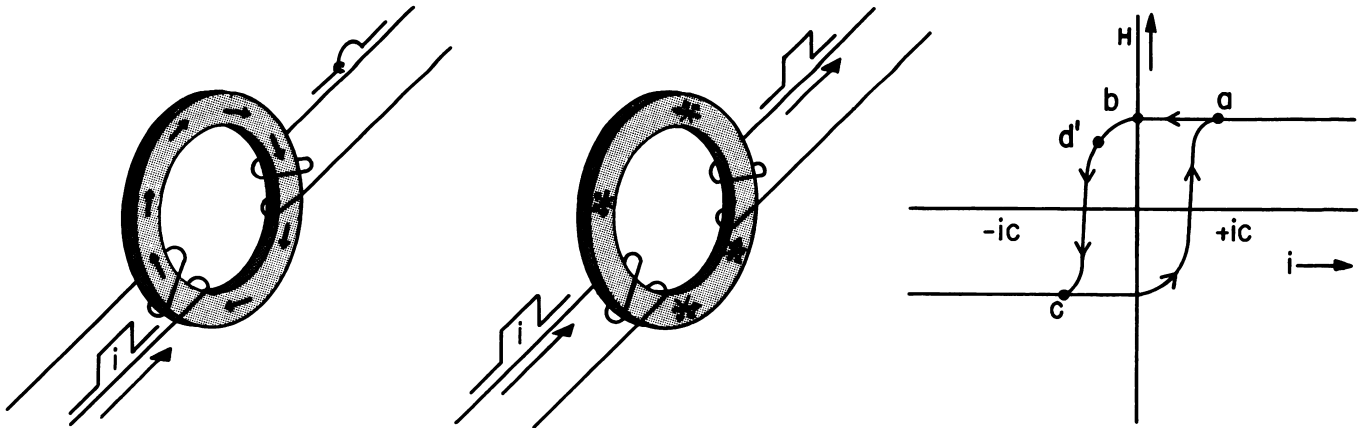
Normal current flow is from +V to -V. A positive voltage input will raise the potential at x, diverting the current to the transistor base. Since the base current is independent of the size and number of input signals, many input lines may be connected to a single gate.

CURRENT MODE LOGIC



Transistor Q₁ is normally conducting, being cut off when any other transistor is conducting sufficient current to raise the voltage at x. Transistors are biased to prevent hole storage problems, thereby permitting very high speeds. Both normal and inverted outputs are available.

magnetic cores



Another device, used mostly as a memory element, is the magnetic core. The core consists of a ferrite toroid with at least two windings. A positive pulse through the input winding will create a magnetic field in the core, in a clockwise direction. This corresponds to point A on the hysteresis curve. When the pulse is removed, the magnetic field will remain (point B). The core "remembers" that it has been set. If a negative pulse is sent through the input winding, the magnetic field will be reversed (point C). The rapid reversal of the magnetic field will induce a current in the output winding. The pulse which sets the core is called the write pulse. A pulse in one direction will write 1; a pulse in the other direction (or a pulse through a separate winding wound in the opposite direction) will write 0. Once the core is set it can be "interrogated" by sending a pulse in the same direction as is used to write 0, called a "read" pulse. If the core is set to 1, there will be a pulse induced in the output line. If it is set to 0, there will be no output pulse. Note that readout is destructive. The process of interrogating the core resets it to 0, regardless of the information content.

Ferrite cores can also be used for logic circuits. The ferrite material is chosen because of its square hysteresis loop. There is a definite minimum current (I_C) required to cause the magnetic field to change direction; a lesser current will not affect the core. If

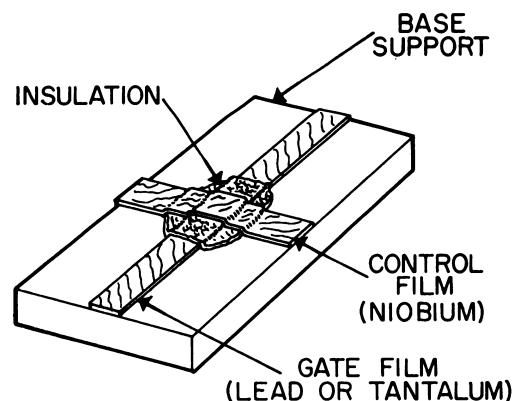
two input windings are used, and a 1 is represented by a current of $0.6I_C$ in each winding, the core will be set only if both inputs are 1. Therefore, it is an AND gate. An OR gate can be constructed by arranging the winding so that either pulse will set the core. If this arrangement is used, it is necessary to reset the core after each output pulse.

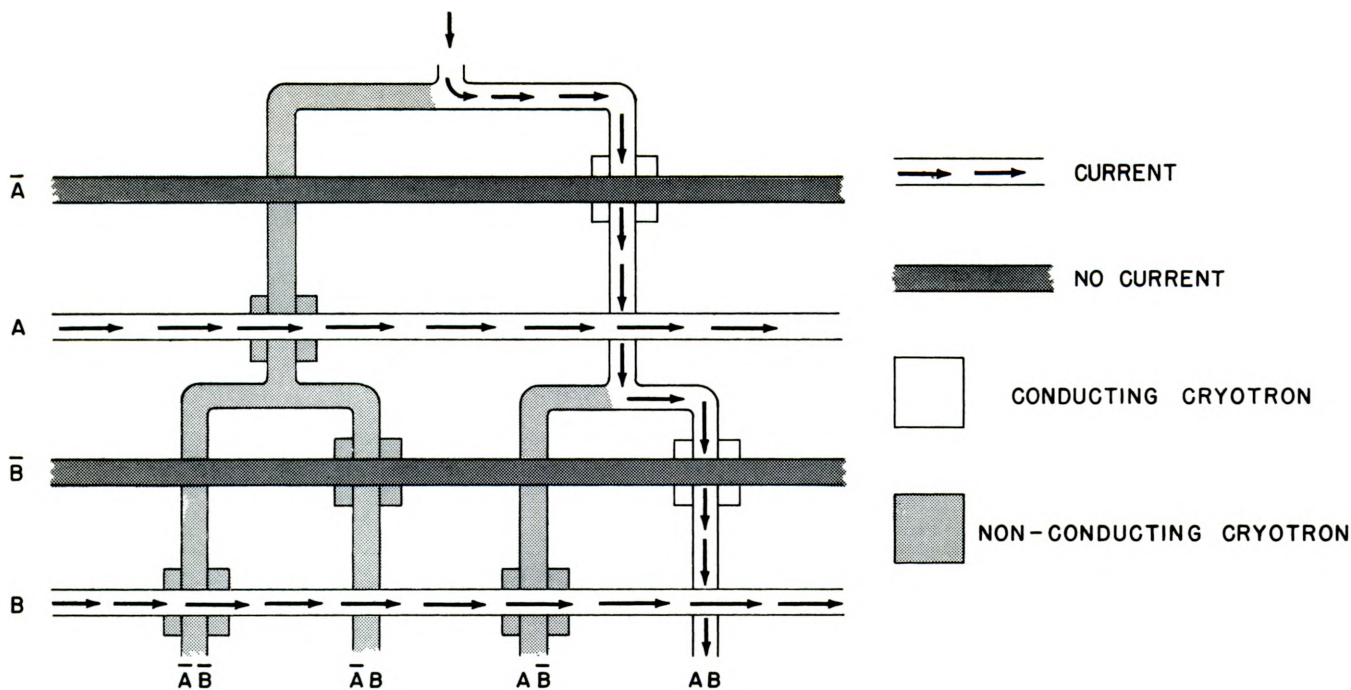
A method of eliminating this problem, and of achieving non destruction readout, is available. It is based on the fact that changing the direction of the magnetic field requires not only a minimum current (I_C), but also that the current be present for a minimum amount of time. If a pulse of very short duration is used, it will end before the field is reversed (point D or D' on the hysteresis curve). When the pulse is removed, the core will return to its previous state.

The vast majority of digital computer circuits use these four basic elements (relays, diodes, transistors and magnetic cores). Many other devices have been used in experimental computers. Micro-miniature binary devices with extremely short switching times are of particular interest. It is not possible to predict which, if any, of these devices will appear in future operational computers. For example, the cryotron shows promise of significant advantages, but many technical problems must be overcome.

CRYOTRONS

The cryotron depends upon the superconductivity of some materials at temperatures near absolute zero. Materials frequently used are tantalum or lead, and niobium. If the temperature is within 8°C of absolute zero, and the magnetic field strength is below a critical value (H_C), these materials offer no resistance to current flow. The value of H_C for niobium is much greater than for tantalum (or lead). A switching element is formed by crossing two films, one of niobium and one of tantalum (or lead). The niobium film is called the control film. As long as there is no current in the control film, the gate film will be superconducting. If a current of sufficient strength is passed through the control film, it will create a magnetic field which will destroy the superconductivity of the gate film. The value of H_C for niobium is sufficiently high so that the superconductivity of the control film is not impaired by the current in the gate film.





A cryotron logic circuit is illustrated. With current in both film A and film B, the current in the gate films can flow only to the terminal marked AB. Each possible combination of A, \bar{A} , B and \bar{B} limits the current to a single path.

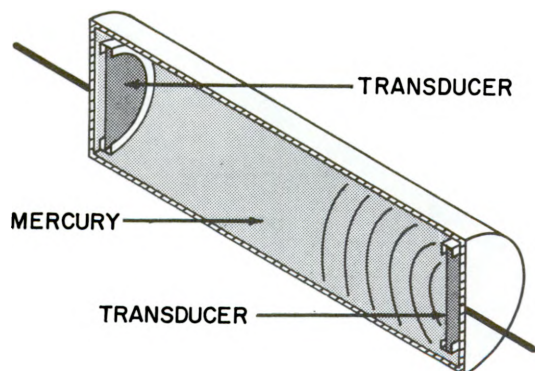
The films may be very thin. In fact, they must be thin, because superconducting materials offer strong resistance to the penetration of magnetic fields. Since a superconductor requires no power to maintain current (once started, a current will continue to flow indefinitely, unless deliberately stopped), the power requirements of cryotron logic is very small. The devices

can be packed into a small volume, since a switching element consists simply of a pair of crossed conductors. The problem, of course, is to develop a cooling system which does not take up more space and power than is saved by miniature design of the logic.

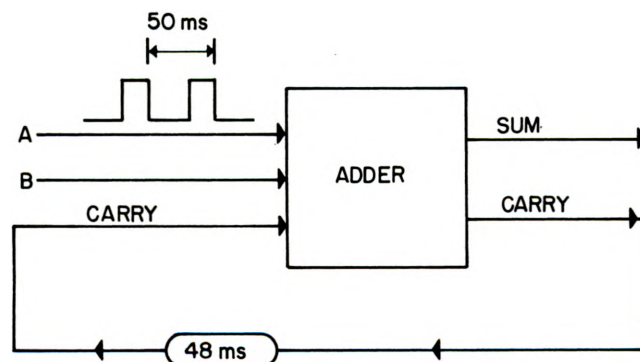
The logic elements described are capable of mechanizing any logical function involving simultaneous events. The AND gate, for example, works only when both inputs occur simultaneously. Devices which can handle sequential events are also required. The two generally used are the delay line and the flip-flop.

delay lines

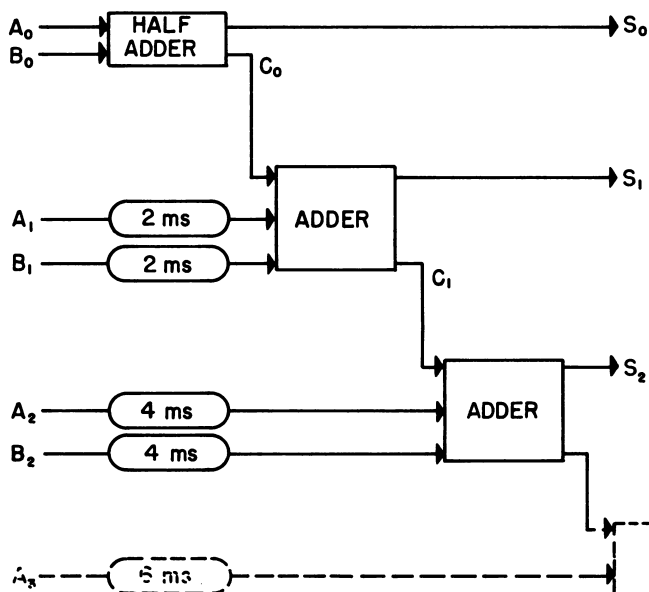
The purpose of a delay line is simply to delay an event a fixed amount of time. The most common form is the acoustic delay line. It consists of two transducers connected by a column of mercury. The first transducer converts an electric pulse to a mechanical vibration, which is transmitted through the mercury and reconverted to an electric pulse by the second transducer. The delay time is equal to the length of the mercury column divided by the speed of sound in mercury.



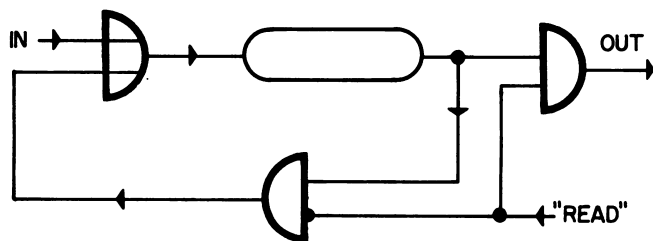
A delay line may be used in conjunction with an adder, to add numbers of several bit length. Assume the two numbers to be added (A plus B) occur in binary form, with each bit 50 milliseconds apart (least significant bit first). Assume it takes 2 milliseconds for a signal to pass through the adder. A 48-millisecond delay line is required to delay the "carry" bit until the next more significant bit occurs. This is called a serial adder. Each number appears on a single line, one bit following another.



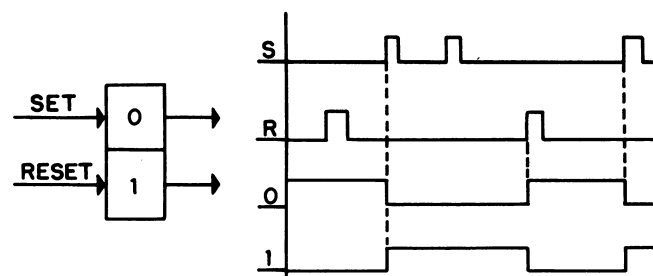
Another possibility is to use a separate adder for each pair of bits. The least significant bit of A and of B is added by the first adder, the second bit of A and of B, plus the "carry" from the first adder, are added by a second adder, and so on, up to the capacity of the machine. In this case, since all bits occur simultaneously, it is necessary to delay the input. The second input must be delayed 2 milliseconds to await the "carry" from the first addition, the input to the third adder must be delayed 4 milliseconds to await the "carry" from the second, etc. This is called a parallel adder.



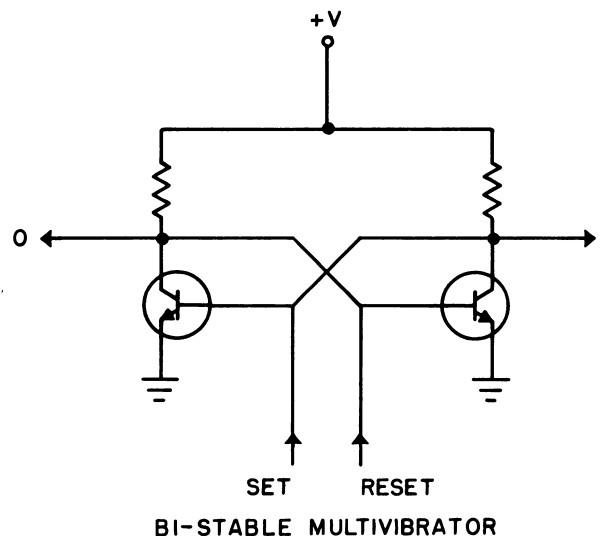
The delay line can also be used as a storage device. If the delay time is greater than the time between bits, several pulses will be passing through the delay line at one time. For example, if the delay time is 200 milliseconds, and bits occur every 10 milliseconds, 20 bits will be contained in the delay line at one time. The arrangement of gates illustrated will keep the data circulating in the delay line until a "read" command occurs.



flip flops



The flip-flop is a device with two input and two output lines. In operation, one output or the other is 1, but never both. Initially, when the flip-flop is in the "reset" or "zero" state, the 1 output is from the "0" side. When a 1 input occurs at the "set" terminal, the device "flips", and the 1 output is from the "1" side. The flip-flop is then in the "one" or "set" state. It will remain in this state until a 1 input occurs at the "reset" terminal. This will cause the device to "flop", i.e., return to the "reset" or "zero" state. When the flip-flop has been set, no subsequent input at the "set" terminal will affect it, nor will it be affected by a "reset" pulse when in the "zero" state. A slightly modified flip-flop (called a complementing flip-flop) is also used. This device has only one input line, and changes state every time a 1 input occurs. The flip-flop can be used as a memory device, since it remains set after the input pulse is removed. Flip-flops are usually electronic bi-stable multivibrators, although a pair of ferrite cores or relays can be used.

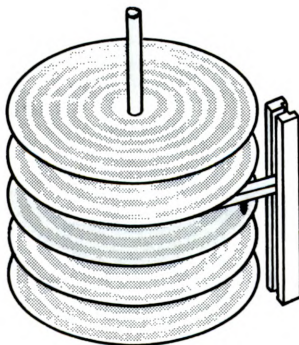
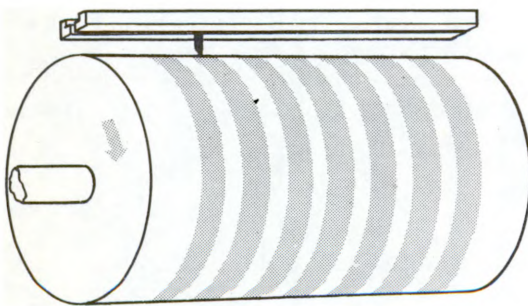


The various functional units which make up a digital computer, viz, memory, arithmetic, control and input/output, were introduced earlier. The manner in which the various digital circuits just described are used to mechanize these functions is described here.

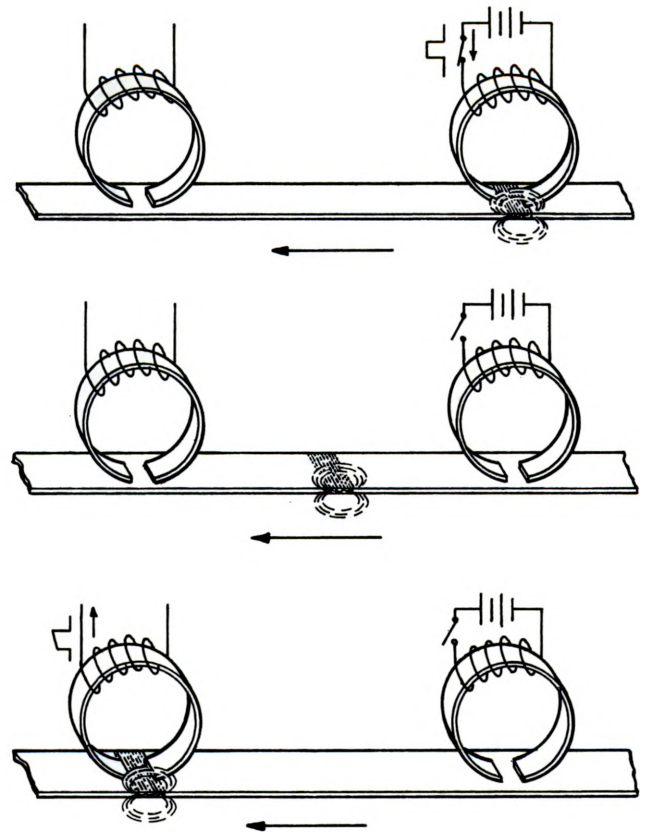
MEMORIES

The function of a computer memory is to store the initial data, instructions, and intermediate results, so they are accessible to the computer at all times. Many different devices are available, and one computer often uses more than one type.

Many computer memories are based on the familiar magnetic tape recording principle. When a strip of plastic, coated with iron oxide or some similar magnetic material, is passed close to the field of an electromagnet, the particles are oriented or polarized. The orientation is retained after the field which created it is removed, and the particles themselves form a miniature magnet. When this magnetized section of tape passes a gap similar to the one which generated it, it will create a magnetic flux whereby it can be read. In practice, several channels are recorded across the width of the tape. The form of the memory is sometimes a tape recorder, but it is often in the form of a drum or a series of disks. There is no basic difference in principle between these forms; the drum can be thought of as a series of tape bands, and the disks as a series of tape rings.

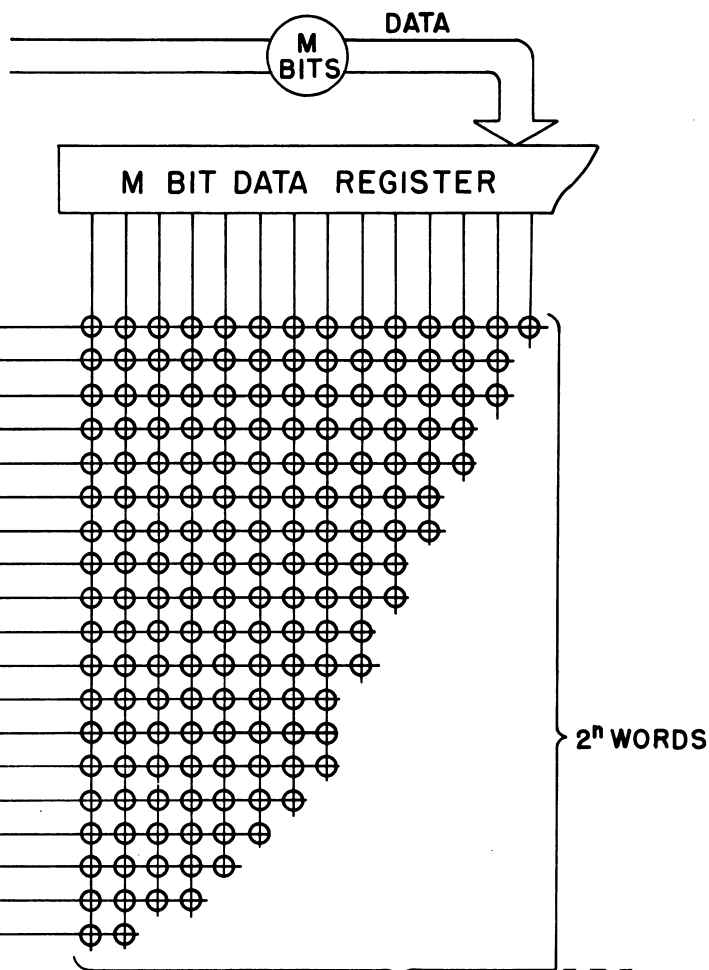


Drum memories and disk memories may read and write one bit at a time, or have several heads in parallel, as is the practice with tape. Some drum memories and disk memories have two or three sets of read/write heads. While one word is being read, another head moves to the address of the next word.



The tape memory has an infinite capacity if allowance is made for changing tapes, and has the largest capacity of all memory devices, even if only the tape actually on the machine is counted. Unfortunately, the access time, i.e., the time it takes to move the tape until the word addressed is under the read head, may be very long, since the entire length of the tape may have to be traversed. For this reason, tape is seldom used for storing intermediate data, although it is often used for storing instructions, since the machine usually runs through instructions, one at a time, in the order in which they occur. The drum-type improves the access time at the expense of capacity. The disk-type also sacrifices capacity for improved access time. The capacity of the disk-type is greater than that of the drum-type, and its access time is about the same if several read/write heads are used, but its cost is greater. The access time for drum-type and disk-type memories can be reduced by optimum coding. This is the practice of storing the instructions in such a way that the material addressed will come under the read head just at the time it is required. This is not practical for general purpose computers. Since the time required for each computer operation must be calculated to determine what position the read head will be in, the process is costly and usually requires several trial runs on the computer. For a special purpose computer, however, optimum coding can often permit the use of a much less expensive memory than would otherwise be required.

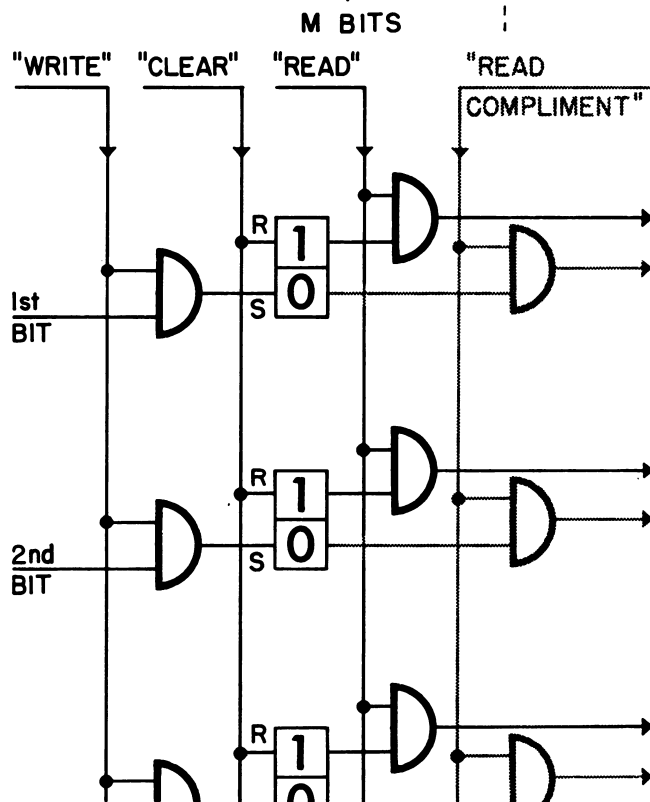
A random access memory, i.e., one in which every word is available at all times, is usually required by general purpose and very high speed special purpose computers. The most common random access memory is the coincident current magnetic core type described earlier. Another type uses small ceramic beads which acquire and hold a static charge when a voltage is applied to them. Both types can be packed densely, and read in a few microseconds, but are very expensive. A "word organized" magnetic core memory is arranged as shown, where each circle represents a core. Each horizontal row contains a word, and each vertical column contains a bit. The decoded address causes a $0.6 I_c$ current in the horizontal line addressed, where I_c is the amount of current required to switch a core. Simultaneously, each "1" bit of the word to be stored causes a $0.6 I_c$ current in the vertical line corresponding to the bit. A one is stored in each core where these currents coincide.



All these are non-volatile memories. The information will be retained even when power is removed, and no power is consumed except when information is being transferred.

For strictly temporary storage, particularly in buffers, flip-flop memories are often employed. A write command permits each bit of the word to be stored to set a flip-flop. The read command resets all flip-flops to the zero state. Since a flip-flop and two gates are required for each bit, and there are usually several components to each of these elements, this type of memory is bulky and expensive. Furthermore, considerable power is expended, even when data is not being transferred, and the memory is volatile. There are several compensating advantages. Readout is very fast. If separate "read" lines are provided, any part of a word can be read, and in any order. With separate "write" lines, bits can be entered in any order. By connecting output lines to the zero side of the flip-flops, as well as to the one side, the one's complement of a word can be read directly if desired. These advantages are of little value in permanent storage, but are helpful for data which is actually in use. This type of memory is frequently used to hold the total in an accumulator, to hold instructions in a "current instruction register" while the operations called for are set up by the control unit, and to hold instructions and addresses being decoded.

Special purpose computers sometimes use permanently wired random access memories. These can be read, but have no provision for writing. The data is implanted when the unit is built, and remains unless the memory is removed and components are replaced or altered.



ARITHMETIC

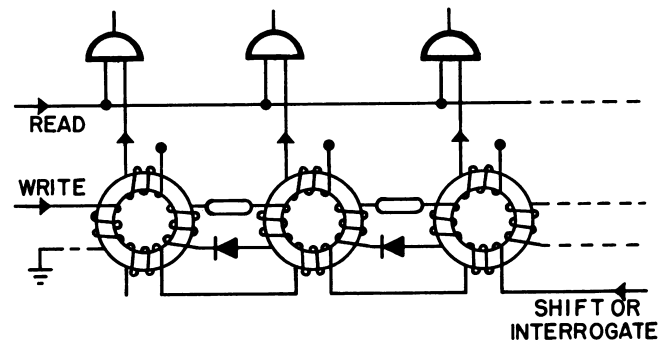
The construction of an adder was discussed in detail earlier. The fact that subtraction, multiplication and division can be performed by adders was demonstrated. The simple adder is capable of adding only two bits which occur simultaneously. This is not convenient if several numbers must be added. To add four numbers (A, B, C, D) three separate additions would be required:

1. Add A and B, store in location Z
2. Add C and D, store in location Y
3. Add the contents of location Z to the contents of location Y.

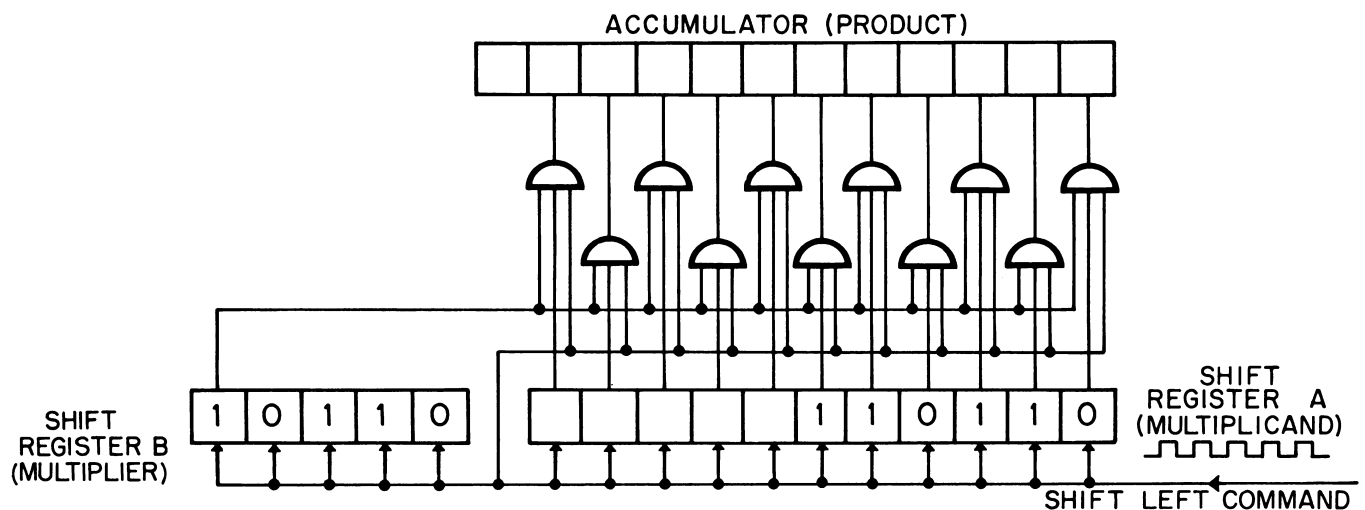
An accumulator is used to simplify this procedure. This is a device which stores the first input, and then adds each successive input to the current total, and stores the new total until it is cleared. To add four numbers using an accumulator, the procedure would be simply to transfer A, B, C and D to the accumulator. This would result in the total of A plus B plus C plus D being stored in the accumulator.

The arithmetic unit also requires shift registers in order to multiply and divide. Part of a magnetic core

shift register is illustrated. Note that each core is connected to the next core to the right. A "shift command" pulse applied to each core will cause all cores to be reset, producing an output pulse from each core which is set to 1. This output pulse occurs not only on the "readout" line, but also on the line to the adjacent core. This will cause the next core to be set to 1 after a delay. Therefore, each "shift command" causes each bit to be shifted to the right.

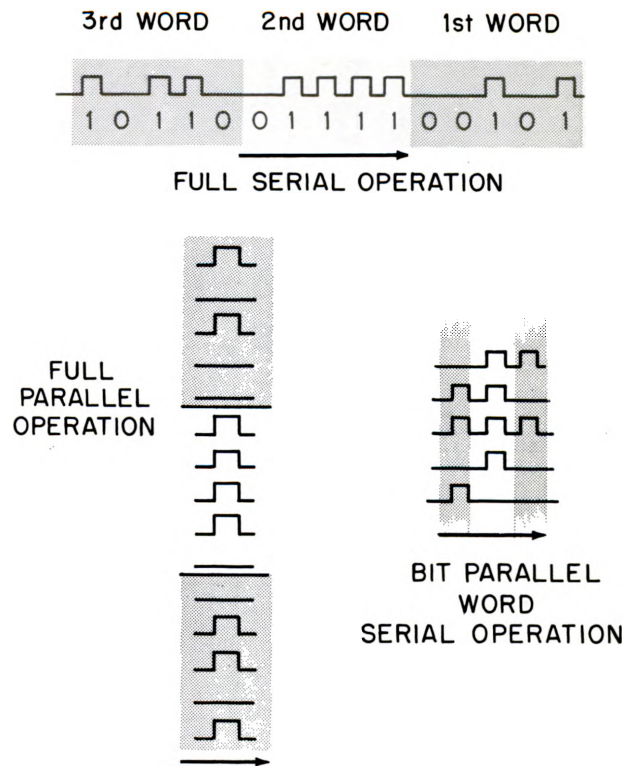


An example of the application of the shift register and accumulator is the method of multiplication illustrated. The multiplicand (110110 in this case) is read into shift register A, and the multiplier (10110) read into shift register B. When the first "shift command" occurs, the multiplicand will be transferred through the AND gates to the accumulator (the number read out of a shift register is that which existed before the shift). Note that the AND gates will transfer the contents of shift register A only when the left hand bit in shift register B is 1. On each succeeding shift pulse, the shifted multiplicand will be added to the total in the accumulator if the corresponding bit of the shifted multiplier is 1. Therefore, to multiply a five-bit number, it is only necessary to provide the five "shift command" pulses to the shift register. These commands are supplied by the control unit.

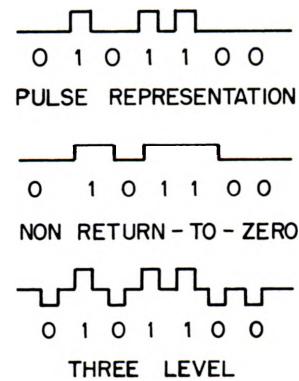


CONTROL

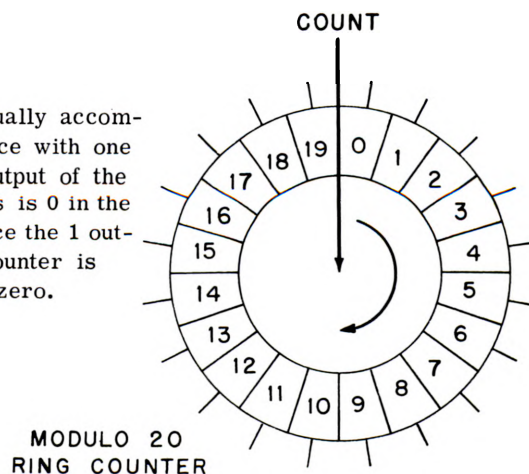
The most critical function of the control unit is its timing of the various operations. To accomplish this, the control unit has some form of a clock. The time scale used depends on the type of data transfer and the switching speed of the logic elements. Many computers accept data in serial form, i.e., one bit at a time. Large, high-speed computers accept an entire word in parallel. The most common format is a compromise whereby parts of the word are accepted serially, but all bits of each part are accepted in parallel. Computers operating on the binary coded decimal system may operate on a bit parallel, digit serial method. The fundamental unit of time is called bit time. This is the time between the occurrence of each bit. Usually the clock generates one pulse per bit time, which is carefully shaped to optimum dimensions. These are called clock pulses. Other times are measured as multiples or submultiples of one bit time. For a BCD system machine, for example, "digit time" would be four bit times. The relative speed of a digital computer is usually expressed in terms of the frequency of the clock pulses. For example, a computer which produces one clock pulse every millisecond is said to have a 1 kilocycle clock rate.



Computers usually operate on pulses or a non-return-to-zero system of representing bits. In either case, the timing is the only positive method of distinguishing between bits. The pulse method gives a positive indication of the presence of a 1, but the only difference between one 0 and twenty 0's is the timing. The three-level system provides indication of the presence of both 1's and 0's, but this is basically a trinary system, and requires more complex circuitry. The accurate indication of each bit time is essential. The confusion which would result if the timing were off by one bit time, and the computer interpreted the last bit of one word as the first bit of the next word is easily imagined.

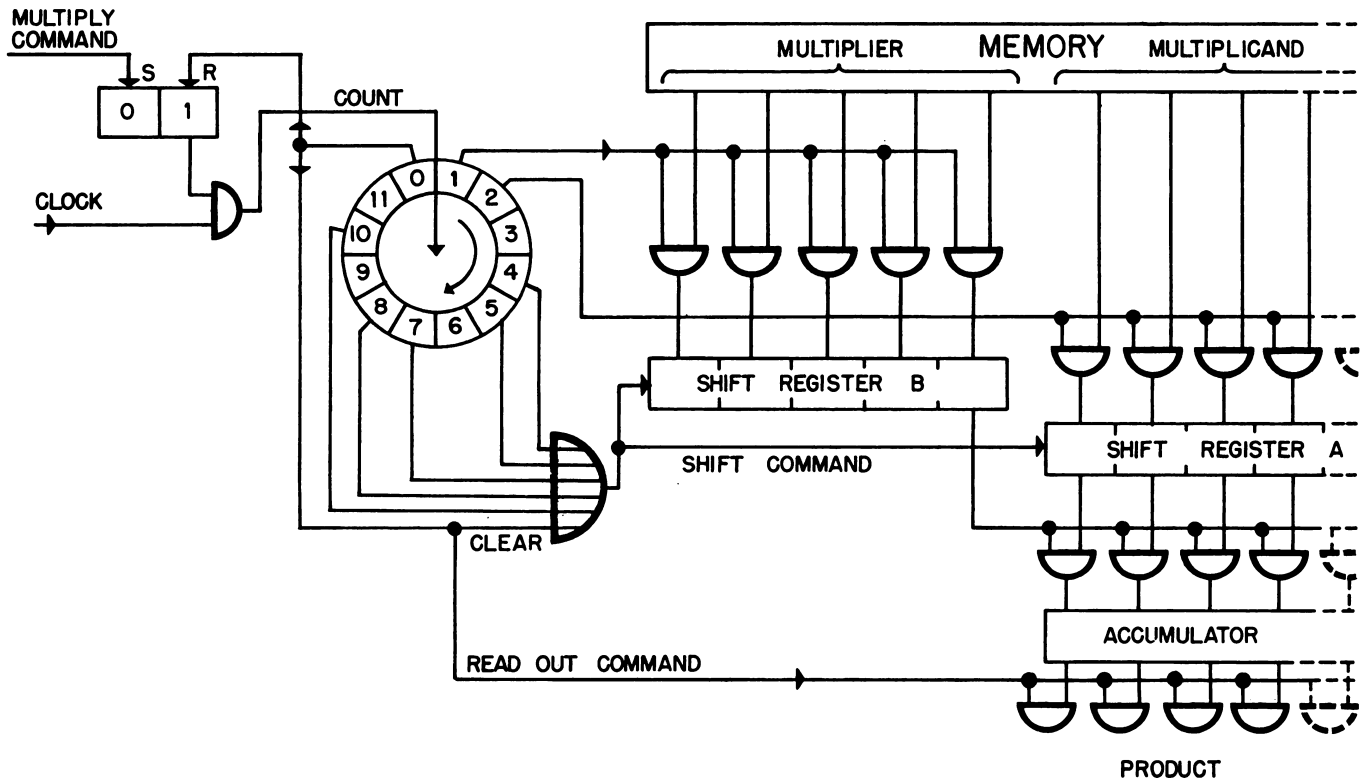


The timing of arithmetic operations is usually accomplished by a ring counter. This is a device with one input line and several output lines. The output of the 0 line is 1, and the output of all other lines is 0 in the normal state. Each input pulse will advance the 1 output one position until the capacity of the counter is reached; then, the counter will return to zero.



To control multiplication using the shift registers and accumulator described earlier, for example, a modulo twelve ring counter might be used. The first output would be connected to gates which transfer the multiplier to the B register, the second output would transfer the multiplicand to the A register, the third, fourth, sixth, eighth and tenth outputs generate five "shift commands", the twelfth would reset the counter to zero and signal the end of the operation. The counts omitted will give the accumulator sufficient time to complete each addition before the next step occurs.

Upon receiving the multiply command, clock pulses would be transmitted to the counter until the end of the operation causes the flip-flop to be reset. Another method of controlling operations employs the preset ring down-counter. This device is much like the normal ring counter, except that it is preset to a number and then counts down to zero. With this device, one counter can control an operation even though the operation is not always performed the same number of times. The control input simply presets the counter for the number of operations desired.



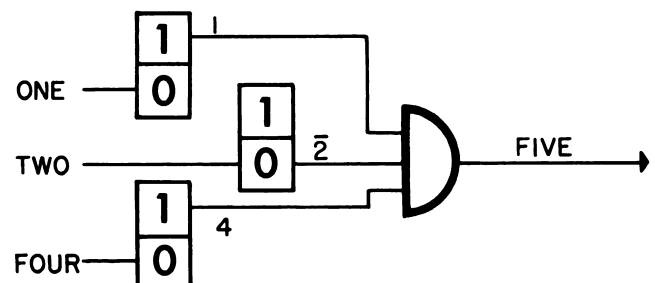
Rather than include a complex arrangement of counters and logic control circuits for operations other than the basic arithmetic functions, general purpose computers sometimes have permanently stored instructions called subroutines. Normally these are stored in a special section of the memory. These might include often used processes such as extracting roots of numbers, forming sines, cosines or exponential functions, etc. If the program requires one of these operations, the control will go to the proper memory locations, execute the instructions called for one by one, and then return to the next step in the main program.

The control unit must also provide the "multiply command" and make the proper numbers available from the memory. These functions are performed by decoding instructions and addresses from the program.

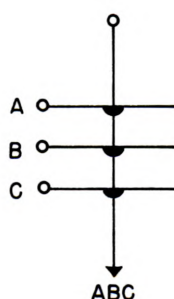
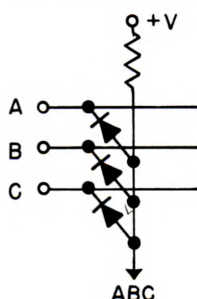
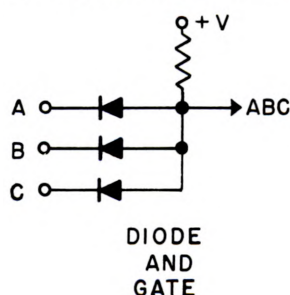
A typical instruction might consist of four words containing the operation to be performed (add, subtract, multiply, divide), the address of the operator (multiplier), the address of the operand (multiplicand), and the address in which the result is to be stored. These

would be stored in the "current instruction register" of the control unit, which may be a flip-flop memory connected to AND gates.

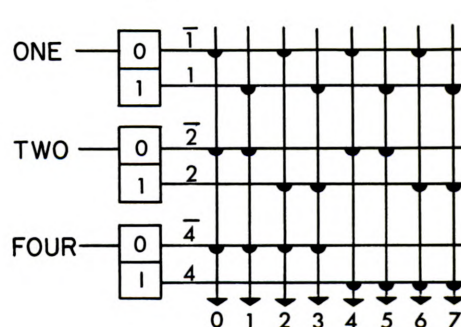
Assume the coded instruction word for multiplication is the number five. Since the binary form for five is 101_2 , the code for five is: four, AND NOT two AND one, or 421. Therefore, the gate for five is connected to the 1 side of the first and third flip-flop, and to the 0 side of the second flip-flop (the output from the 0 side will be 1 if the flip-flop is storing a 0). The output of this gate would be the "multiply command".



A similar arrangement is used to decode the address of the operators. One AND gate is required for each code. Diode AND gates are most often used. The illustration shows how the schematic diagram for an AND gate can be redrawn in a matrix configuration, and how the matrix is diagrammed symbolically. The diagram for a complete three-bit decoder shows that the matrix



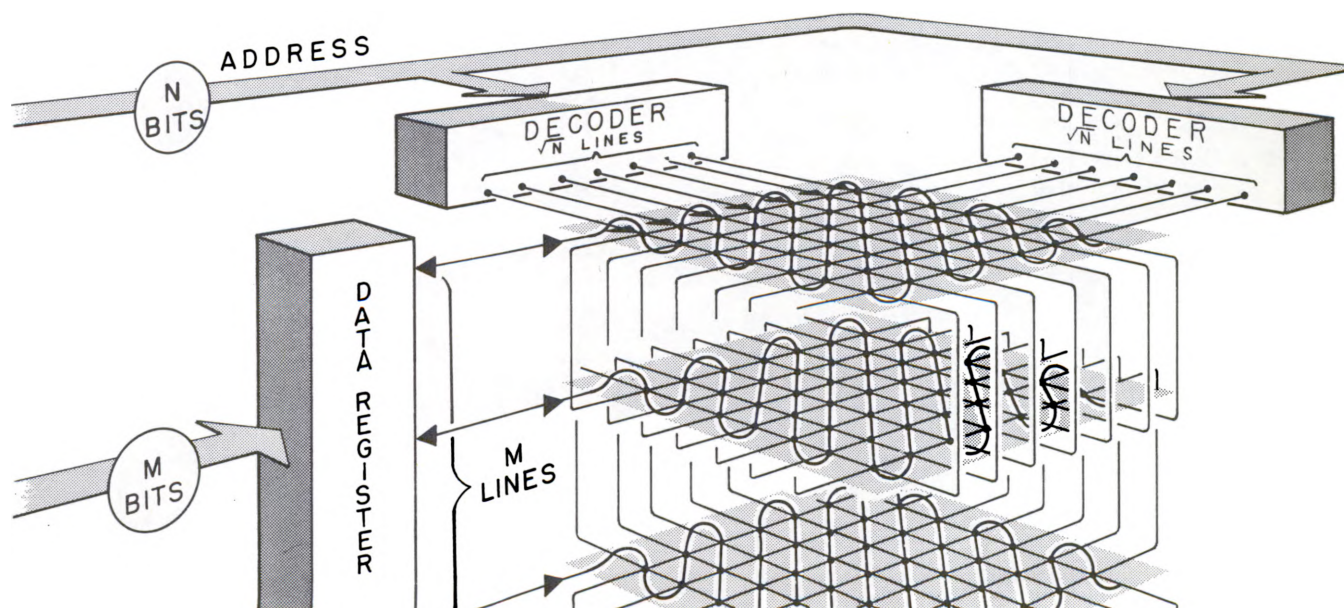
MATRIX FORM



Examining the matrix for the decoder shows that the number of diodes required is equal to the number of bits in the address times the number of possible addresses (words). Since an n bit address can code 2^n addresses, the number of diodes required is $n \times 2^n$. Computers often have 5,000 to 15,000 addressable memory locations. A 4,096-word memory requires a 12-bit address and $(12 \times 4,096)$ 49,152 diodes. A 15,000-word memory using this type decoder would require over a million diodes. Small, cheap diodes are available, but they are not sufficiently small or cheap. For large capacity memories, a three-dimensional or "bit organized" layout is employed. As shown in the illustration, each plane corresponds to a bit. Each word is addressed by a combination of two codes, and each core has three windings. A current of $0.6I_C$ is sent through each of two address lines. This will cause a "1" to be stored in each of the cores for this word (each vertical row is a word). If a 1 is to be stored, no current is sent through the bit line. If a zero is to

is much easier to interpret than a large number of AND gates. Note the cyclical pattern of diode connections along the lines from left to right. The connections to the "one" flip-flop alternate from "1" to "0" on each line, those from the "two" flip-flop on every second line, and those to the "four" flip-flop on every fourth line.

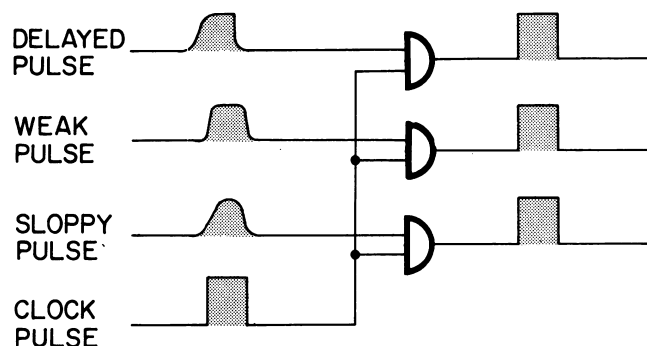
be stored, a current of $-0.6I_C$ is sent through the bit line, reducing the total current below the critical current. In this arrangement, each decoder decodes only $n/2$ bits requiring $n/2 \times 2^{n/2}$ diodes. Two decoders for the 4,096-word memory would require only $(2 \times 12/2 \times \sqrt{4,096})$ 768 diodes instead of almost 50,000. This large reduction occurs because the cores participate in the decoding. Each core, by responding to a single combination of three currents, is acting as an AND gate, thereby reducing the logic requirement of the decoders. No special read signal is necessary to read out data. A $-0.6I_C$ current from the two decoders will cause current in the bit line for each core in the word which is storing a 1. Further reduction in the complexity of the decoder is achieved by decoding in sections and then decoding the results, rather than decoding an entire address at once. Decoding instructions presents much less of a problem since a computer will rarely operate on more than a few basic instructions.



There are addresses other than those of the main memory, of course. As a rule, the accumulator, the various shift registers, instruction registers, etc., are also addressable. In order to operate properly, the computer must be able to transfer data at will. Since it is not practical to have every storage location connected to every other location, a routing system is necessary. The memory devices just described, for example, require that the data to be stored be in the particular register to which the cores are wired. If a number is to be transferred from a main memory location to the accumulator, it must be transferred first to the memory data register. This routing is sometimes handled by the program. In this case, the programmer must first instruct the computer to transfer the data from the main memory location to the memory data register, and then to the accumulator. In more complex machines, the routing is accomplished by the control unit.

One incidental use of clock pulses is reshaping signals. As the signal pulses pass through various gates, they

may become distorted or delayed. To correct these conditions, the signals are gated with clock pulses at various points throughout the computer. Although an AND gate is used, it serves as a switch and has no logic function. Since a clock pulse is present during each bit time, the output is logically identical to the signal input, and the gate shapes the signal in conformance with the clock pulse configuration.



time - sharing

The most complex control function is that of handling more than one problem at a time. One method of accomplishing this is time-sharing. This is a procedure often used with special purpose computers which must handle a few problems over and over. The situation arises often in weapons systems, whereby two or more fairly similar problems must be handled simultaneously, or where the same operations must be performed on several distinct sets of inputs.

Such a situation might arise in a computer which directs surveillance radars, displays targets encountered, and predicts their future courses. Here, one computer might assign different search patterns to several radar sets, and accept data from all of them. The computer would then select an individual target, compute its position and velocity, predict its future course, assign an identifying code, store and display the data, then select a second target and repeat the procedure, continuing this process until all detected targets are displayed.

When all targets are accounted for, the cycle would begin again with the previous data corresponding to the target under consideration recalled from the memory. The control is not very complex for this type of operation, because the program is the same for all targets, and at any instant the computer is handling only one problem.

Another type of time-sharing occurs when a special purpose computer which would normally be idle a large part of the time, handles less essential, but more time-consuming problems during idle hours. When the computer is needed for its primary function, processing of the low priority problem will be halted, and resumed when the high priority problem is completed. There is no great increase in complexity in this type computer, as again, only one problem is in progress at any instant.

In other applications, when a fairly complex general purpose computer is handling problems which do not require its full capacity, the programmer may start

a second problem, using portions of the computer not required for the first problem. In this case, it is the responsibility of the programmer to prevent conflicts. The most sophisticated control circuits are found in those computers which can handle two or more programs at the same time, even though the same computer components are used for both problems. In this case, the computer must keep each program separate. Before beginning a step, the computer must check that the components required are not in use by another program. If the computer cannot proceed with the step, it will wait until the data or arithmetic circuit is free. This type computer normally uses a "forwarding address" system. Here the programmer assigns memory locations for the data and instructions, just as in a normal computer. Since the same address can be assigned to different data by different programs, the computer separates the addresses according to program, and then assigns the data to any vacant location. Therefore, if program A assigns data to memory location 108, the computer will place the data in an empty memory location, and place the address of this location at address 108, along with an identifying code showing that this location applies to program A. If program B also assigns data to location 108, a different "forwarding address" will also be stored in location 108 with an identifying code for program B. When a program calls for reading data from location 108, the computer will go to location 108, determine the appropriate "forwarding address", and read the data from that address.

This reduces the memory capacity, of course, since some locations must be reserved for forwarding addresses, and others reserved for data. This is part of the price which must be paid for the ability to handle separate programs simultaneously. Obviously, the complex control unit necessary for this type of operation is very expensive. In order for the increased benefits to be worth the price, the computer must be extremely complex.

INPUT / OUTPUT

The variety of devices used to provide input data to a computer and produce output data in a useable form is too numerous for detailed coverage here. These devices perform several separate functions, viz., form conversion, reading and writing, buffering and data conversion.

FORM CONVERSION

Form conversion, i.e., changing voltage levels or waveshape, is the simplest function of input-output equipment. When the input data occurs in digital form and in the same code used by the computer, the input equipment may simply adjust the input data to the proper voltage and waveshape, or possibly convert from a non-return-to-zero to a pulse mode, or vice versa.

READING AND WRITING

A reader converts some physical data presentation, i.e., magnetic impressions on tape, perforations in tape punched cards, cathode-ray tube display, typed or printed pages, photographic images, depressed keys, etc., into electrical pulses suitable for computer operations. A writing device reproduces output data from the computer into one of these physical presentations.

BUFFERING

A buffer is a device which adapts the rate of flow of a reading or writing device to the rate of flow of the computer. A magnetic tape reader may operate at the rate of 1,000 frames per second, and contain 8 bits per frame, for a total average rate of 8,000 bits per second. The computer may accept data at this rate, but one bit at a time. In this case, a buffer would accept 8 bits in parallel, at the rate of 1,000 bits per second, and allow them to be read out serially at the rate of 8,000 bits per second by the computer. Or, the computer may produce output data at a low average rate, since results may take a long period of time to obtain; but during the period when the results are available, the output rate may be temporarily very high. In this case, a buffer would store many computer words, and permit the results to be printed at the relatively slow rate of a line-printer or automatic typewriter. Since a considerable length of time may pass before the computer completes another calculation, all the data can be read out by the time a new output from the computer occurs.

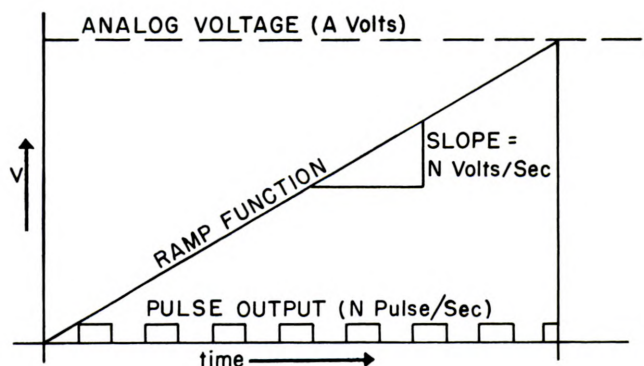
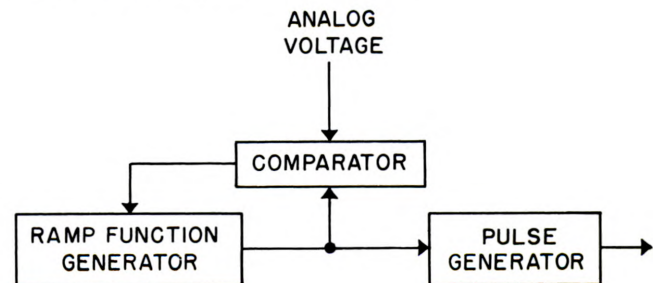
DATA CONVERSION

Data conversion equipment converts digital data from one code to another, or converts from digital-to-analog or analog-to-digital form. Code conversion is usually accomplished by a decoder similar to those described in the control unit discussion, plus an encoder which operates on similar principles.

Analog-to-digital converters occur in many forms, and are used internally as well as in input/output equipment. Two types (one for voltage, and one for shaft position) are used frequently enough to warrant discussion here.

The time-frequency voltage analog-to-digital converter generates a ramp function with constant slope which is cut off when the amplitude is equal to the amplitude of the analog voltage. The ramp function

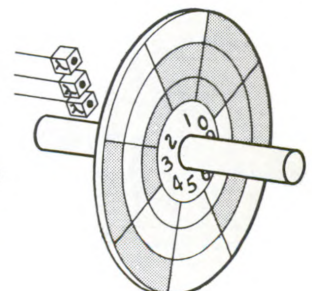
drives a fixed frequency pulse generator. Since the number of pulses is proportional to the period of the ramp function, and the period of the ramp function is proportional to the amplitude of the analog voltage, counting the pulses gives a digital output proportional to the analog voltage. Converters can sample an analog input as often as 1,000 times per second, and produce a digital output accurate to 0.1 percent.



$$t = \frac{A \text{ Volt}}{N \text{ Volts/Sec}} = \frac{A}{N} \text{ Sec}$$

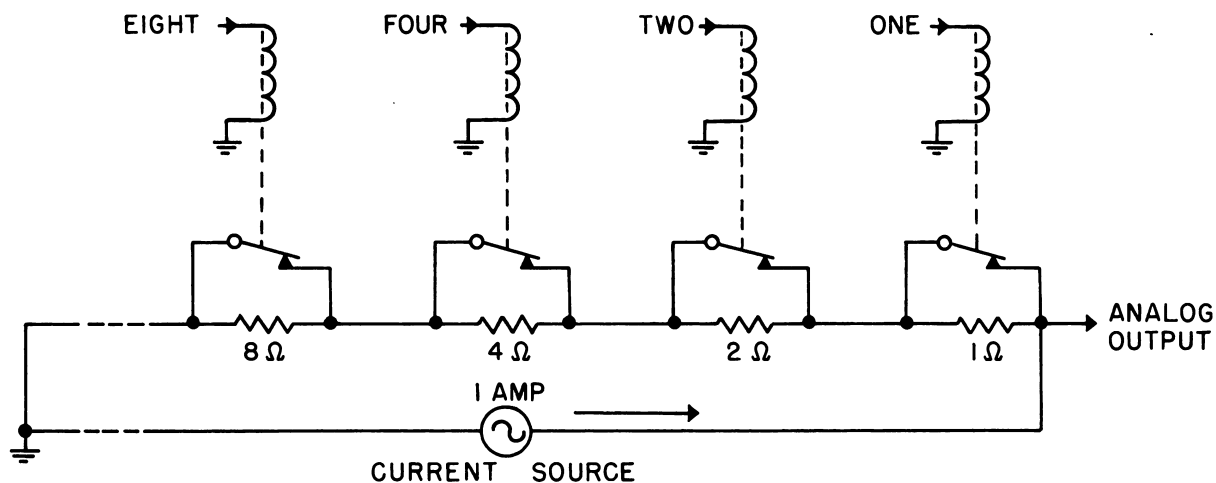
$$N \frac{\text{Pulse}}{\text{Sec}} \times \frac{A}{N} \text{ Sec} = A \text{ Pulses}$$

A coded disk shaft-position analog-to-digital converter picks off a code number from a disk with brushes or photo cells. The cyclically-permuted or Gray code is generally used in this application. If the sensors happen to fall right on the dividing line between two sectors, the device tends to read a 1 if it occurs in either sector. This can cause a gross error if the binary code is used. For example, if the sensor falls on the boundary between sector eight (01000) and sector seven (00111), it will usually read fifteen (01111). The same situation in a cyclically-permuted code will result in reading an eight. Since each sector differs in only one position from the preceding sector when the cyclically-permuted code is used, the device will always be read as one of the two possible sectors for an average error of 1/2.

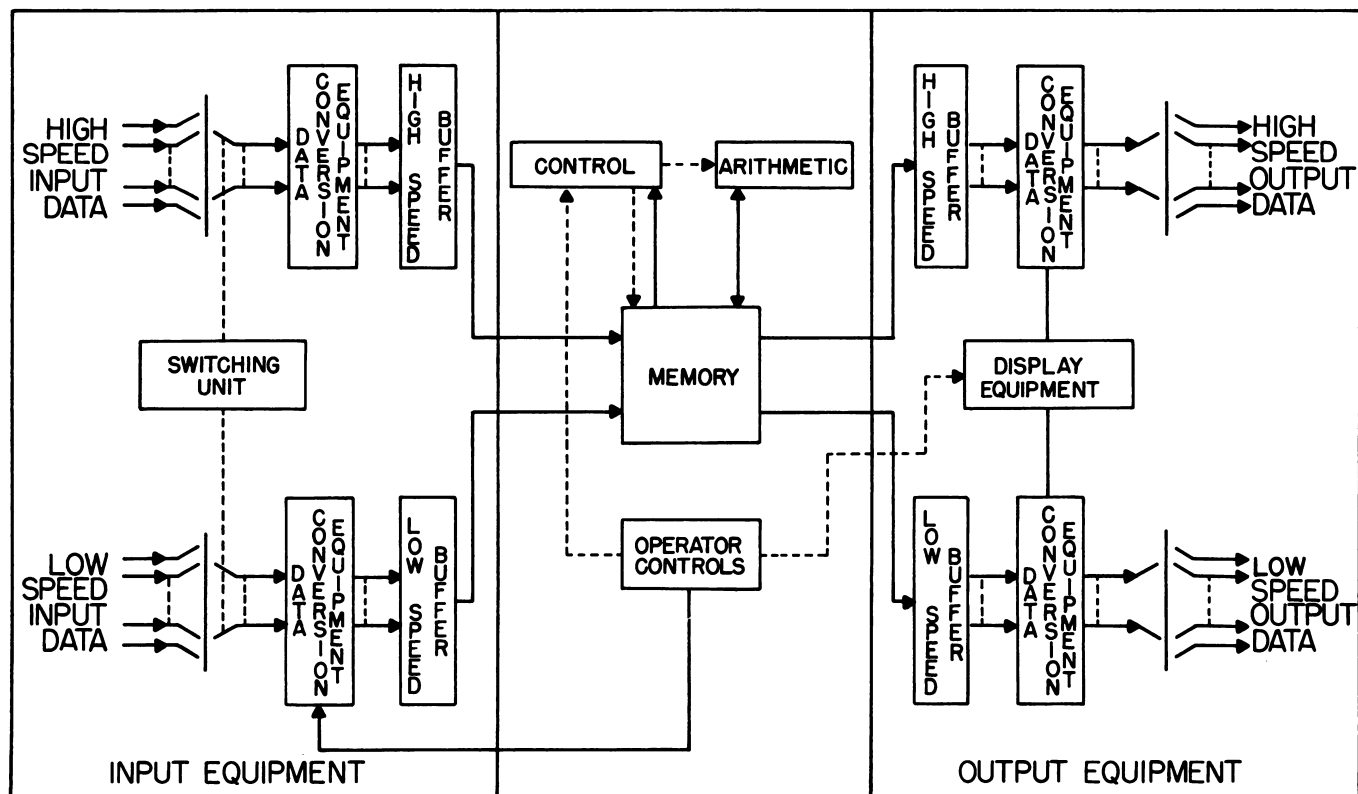


Electrical digital-to-analog converters usually operate by adding voltages equal to the value of each bit position. The converter illustrated will create an output voltage numerically equal to the binary coded input by

adding the appropriate voltages. Note that each binary 1 causes a voltage equal to the value of the position, i.e., 1, 2, 4, 8, 16, to be added to the output.



The relation of these various functional units in a typical military digital computer is illustrated below:



OPERATION OF DIGITAL COMPUTERS

The digital computer uses methods essentially the same as a human performing calculations with pencil and paper. The computer, however, is much faster, much more precise, and very stupid. Because of this stupidity, the program must contain precise instructions for each step, together with what decisions must be made and what criteria must be applied for these decisions. Every conceivable contingency must be provided for. The machine does not know what it is doing, and therefore cannot know if a mistake has been made.

One possible source of error is overflow. If two large numbers are added or multiplied, and the total exceeds the capacity of the machine, the most significant bits will be lost. The machine will simply continue with the data which remains. To avoid this error (and many similar errors), it is necessary either to provide a warning when the error occurs, build a machine capable

of avoiding the error, or take sufficient care in programming so that the error does not occur. Providing a warning is not efficient, since it requires stopping the machine and checking the program, and the overflow does not necessarily indicate an error, since it may be deliberate. The floating-point machine greatly reduces the problem of overflow. This type of machine records numbers with a power of two as a factor. Just as decimal numbers can be written in a variety of forms by multiplying by a power of ten ($986,000 = .986 \times 10^6 = 10^5$, etc.), the binary point in a binary number can be shifted by multiplying by a power of two ($1011000. = 1011_2 \times 2^3 = .1011_2 \times 2^7$, etc.).

Only six bits in the power of two position are required to accommodate numbers from approximately 10^{-10} to 10^{+10} (2^{-31} to 2^{+31}).

address

As pointed out earlier, there are five pieces of information required for each basic operation: two operands, the operation, the destination of the result, and the next instruction. The instructions do not actually contain the operands or the next instruction; they contain only their address. All the addresses may not be contained in one instruction, and some may be implied. A machine which requires instructions containing all four addresses is called a four-address machine, but these are rare. Since, for the most part, the next instruction is contained in the address one number following the present instruction, this address is usually omitted. Instead, an instruction counter is used which starts at the address of the first instruction (usually 001), and adds 1 as each instruction is executed. This is a normal three-address machine. The address of the next instruction is contained in the next instruction register. This register has an address of its own, and its contents can be altered by the machine. Altering the next instruction changes the course of the computation. The importance of this feature cannot be over-

emphasized. This ability to alter its instructions is the primary process in achieving versatility in a digital computer. Using this ability, the programmer can set up a computer to handle a wide variety of problems. Some computers eliminate the address of the destination, automatically storing the result in the location of the first operand. Still further reduction is achieved by including only the address of one operand in the instruction. In this case, at least two separate instructions are necessary for any arithmetic operation. These one-address and two-address systems are usually found in small special-purpose machines.

programming

To illustrate the techniques employed, and the results which can be achieved, a simple problem will be programmed for a hypothetical digital computer. The hypothetical computer uses a three-address system employing a twenty-eight-bit instruction word, eight bits for each address, and four bits for the operation.

The compliment of operations which the computer can perform is listed along with the meanings of α' , β' , and γ' . (The primed notation indicates "the address of", i.e., α' is the address of the memory location which contains α)

Operation	Code	Meaning of α' , β' and γ'	Explanation
Add	0001	Put $\alpha + \beta$ into γ	
Subtract	0010	Put $\alpha - \beta$ into γ	
Multiply	0011	Put first 40 bits of $\alpha \times \beta$ into γ	Ordinary multiplication drops the last 40 bits of the product. If greater precision is required, the double precision multiply instruction is also given, leaving the product distributed in two memory locations.
Multiply-Double Precision	0100	Put last 40 bits of $\alpha \times \beta$ into γ	
Divide	0101	Put α/β into γ	
Compare-Magnitude	0110	If $ \alpha \geq \beta $, set next instruction register to γ	These are the basic instructions which allow the computer to change its procedure according to the results of the calculation.
Compare-Algebraic	0111	If $\alpha \geq \beta$ set next instruction register to γ	
Shift Left	1000	Shift α, β places to the left and transfer result to γ	
Extract	1010	Transfer to γ , each bit of α which corresponds to a 1 in β	This instruction is used to extract part of a data word. To extract for first four bits of α , for example, β would be 1111000
Interchange	1011	Transfer α to γ	
Read	1100	Transfer the next α words appearing on device γ to β , $\beta+1, \beta+2, \dots, \beta$	Since the machine has several input devices, it is necessary to specify the address of the device.
Write	1101	Transfer the next α words from $\beta, \beta+1, \dots, \beta+\alpha-1$, to device γ	
Post-Mortem	1110	Print out contents of all registers along with the address of the register	This is used to determine what went wrong if the computer stops, or if it continues longer than the problem should require without producing an answer.
Stop	1111		

The hypothetical computer is a floating point machine, and all read-out is nondestructive. The information contained in any device will remain until new information is read into it. When the start button is depressed, the machine will read the first instruction on a particular tape, and set the next instruction register to one. The process of programming consists of four fundamental procedures. First: reducing the problem to a series of basic steps. Second: arranging the steps in order, and assigning the variables. Third: coding the program and data in machine language, and assigning memory locations. The fourth step, necessary for

any complex program, is a trial run or debugging procedure to find any faults in the program. This is usually done by running the program with data for which the correct results are known. The third step, coding, is sometimes considered a separate process, since it is a routine procedure often performed by another machine.

The computer programs to find the roots of equations: $x^2 + 3x + 2 = y, y = 0$ is arrived at as follows: (an analog solution of this equation was described earlier).

The method used is a trial-and-error procedure. Assuming intervals of $\pm .0001$ are sufficiently accurate, a value of x is selected and the value of y computed and stored. Then, $+ .0001$ is subtracted from the first value for x , and y is again computed. If the sign of y changes between the two values of x , a root has been found; if not, the $+ .0001$ is again subtracted and the process is repeated. It is assumed that the fact is known that there are two non-imaginary roots, and that both roots are less

than zero. Since the first value of y (at $x = 0$) is 2, this value can be assumed and the first two steps eliminated. The problem requires twelve instructions and data words, so memory locations 001 through 021 are allocated for the program (the first instruction is not stored). The exact number of instructions would not be known initially, and the space would have to be left blank, until this is determined. All numbers are decimal, and numbers in parentheses are addresses.

INSTRUCTIONS

	Operation	α'	β'	γ'	Remarks
(000)	Read in	21	1	Address of tape reader	Read in 21 instruction and data words.
(001)	Subtract	(012)	(013)	(012)	$X_{i-1} - .0001 = X_i$
(002)	Multiply	(012)	(012)	(014)	x_i^2
(003)	Multiply	(015)	(012)	(016)	$3x_i$
(004)	Add	(014)	(016)	(014)	$x_i^2 + 3x_i$
(005)	Add	(014)	(017)	(016)	$y_i = x_i^2 + 3x_i + 2$
(006)	Compare-algebraic	(018)	(016)	(001)	If $y_i \geq 0$, return to (001).
(007)	Add	(019)	(020)	(020)	Add 1 to root count.
(008)	Print out	(019)	(012)	Address of printer	Print out root.
(009)	Transfer	(021)	-	(006)	Change step 6 for negative values of y .
(010)	Compare	(020)	(017)	(001)	If $2 \geq$ root count return to (001).
(011)	STOP	-	-	-	-

DATA

	Initial Contents	Later Contents
(012)	$+ .0001$	x_i
(013)	$.0001$	$.0001$
(014)	-	x_i^2 and $x_i^2 + 3x_i$ alternately
(015)	3	3
(016)	-	$3x_i, y_i$ alternately
(017)	2	2
(018)	0	0
(019)	1	1
(020)	1	root count + 1
(021)	Compare-algebraic (016) (018) (001)	

x_i is the value of x on the i th iteration. $x_0 = +.0001$, $x_1 = 0$, $x_2 = -.0001$, $x_3 = -.0002$, $x_{100} = -.0099$ etc.

000 This instruction is translated as: read the next 21 words appearing on the tape into memory locations 001, 002, 021.

001 It is not necessary to subtract .0001 from x the first time around, but it is convenient to do it this way. If it is required that the initial value of x_i be 0, the original value of x_i can be read in as $+ .0001$ instead of 0.

002-005 Compute y . Note that the locations 014, 016, are used twice. After $x_i^2 + 3x_i$ is computed, x_i^2 and $3x_i$ are no longer needed so their memory locations are reused. Similarly, after y_i is computed, $x_i^2 + 3x_i$ is no longer required. This conservation of memory space can be very important in long problems.

006 This is the critical step of the procedure. In order that the computer may recognize a root when it reaches it, a logical question is asked. Using the knowledge that the initial values of y_i are positive, the computer compares each value of y_i with 0. When the value of y_i is less than 0, a root has been found in the interval $x_i, x_i + .0001$. If the value of y_i is not less than zero, the computer returns to

the first step. This is called a "loop instruction" whereby the computer repeatedly returns to an earlier instruction until a specified condition is met. If the sign of the first value of y_i were not known, a slightly more complex comparison would be necessary.

007 In order to determine when two roots have been found, the computer counts them as it finds each root.

008 As each root is found, it is printed out.

009 After the first root is found, it is necessary to change the order of comparison in step 006, since the next root will be indicated by change from negative values of y_i to positive values.

010 In order to know when to stop, the computer compares the root count with two. If two roots have not been found, the machine returns to the first step. When two roots have been found, the machine proceeds to instruction 011 which tells it to stop. The program would be coded in binary form. For example, the second step would appear as:

0010	00001100	00001101	00001100
subtract	(012)	(013)	(012)

Several points about the program are worthy of emphasis. Since the second root of the equation is -2.000 , the procedure will require 20,000 iterations ($\frac{2}{.0001}$ repetitions of steps 001 through 006) or about 120,000 steps, yet the program contains only 12 instructions and the entire program uses only 21 memory locations. This indicates the programming economy of iterative routines. To solve the same problem using the quadratic formula

$$(X = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a})$$

would require about 35 instructions and 45 memory locations.

The number of iterations could be greatly reduced by using a subroutine. Instead of starting with intervals of .0001, the program would start with intervals of 0.1. When an appropriate root is found, instead of printing out the result, the program would enter a subroutine which would add .001 until the sign changes again, and then return to the main program. If still greater accuracy is required, another subroutine can be used with intervals of .0000001. Obviously, this permits any degree of precision required without an excessive number of iterations.

weapons system applications

An application more germane to weapons system is the solution of the lead angle prediction for a torpedo tube, handled by analog devices in volume I.

The information available is the target range (R) and bearing (B), own ship speed (DM_o), and the torpedo speed (DM_a). The problem is to compute the torpedo tube train (Bdg'). The torpedo will hit the target

Note that the program is not restricted to this particular equation. By simply changing the data, any quadratic equation can be handled. By changing the initial data in 012, the problem can start anywhere. Changing 013 allows any interval to be selected, etc.

Some factors have not been considered, such as imaginary roots, double roots and roots separated by less than .0001. In any actual problem, these contingencies must be provided for. If either of these contingencies occurred with this program, the computer could not find two roots, and would run until it were shut down.

One method of resolving this difficulty is to use a "branch" instruction. In this application it might be decided that if two non-imaginary roots are not found after a certain number of iterations, the computer should start looking for imaginary roots. Two extra instructions would be required, one following 001, and the other following 009. The first extra instruction would cause the number of iterations to be counted; the second would cause the machine to branch if the number of iterations exceeded a present number. Some computers have auxiliary devices (called "B" boxes) to perform such tasks as counting iterations.

(maybe) if the tangential velocity of the torpedo (DM_{ba}) is equal to the tangential target velocity (DM_{bt}).

Therefore, the principal equation is $DM_{ba} = DM_{bt}$. Since the relative tangential velocity of the target is $RxDB$ and the tangential component of DM_o is $DM_o \cos B$, the actual tangential target velocity is $RxDB - DM_o \cos B$.

Therefore:

$$DM_{ba} = RxDB - DM_o \cos B$$

$$\text{Since } DM_{ba} = DM_a \sin L$$

$$DM_a \sin L = RxDB - DM_o \cos B$$

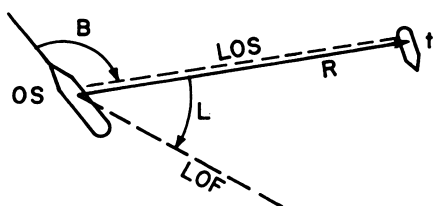
$$L = \sin^{-1} \left(\frac{RxDB - DM_o \cos B}{DM_a} \right)$$

and since

$$Bdg' = B + L$$

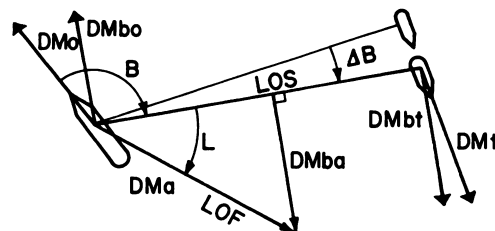
$$Bdg' = B + \sin^{-1} \left(\frac{RxDB - DM_o \cos B}{DM_a} \right)$$

which is the analytic solution of the problem.



GEOMETRY

BEARING	B	LINE OF FIRE	LOF
RANGE	R	LEAD ANGLE	L
LINE OF SIGHT	LOS	TUBE TRAIN Bdg'	$B+L$



DYNAMICS

OWN SHIP SPEED	DM_o	TARGET SPEED	DM_t
COMPONENT OF DM_o TO LOS	DM_{bo}	BEARING RATE ($\frac{\Delta B}{\Delta t}$) DB	
COMPONENT OF DM_t TO LOS	DM_{bt}	TORPEDO SPEED	DM_a
		COMPONENT OF DM_a TO LOS	DM_{ba}

The range rate has been ignored up to this point. The illustration shows three targets, each of which has the same tangential velocity. In the first situation the torpedo will hit the target at point P. In the second situation the impact point is out of range, and in the third situation the target is moving faster than the torpedo and there is no impact point. The third condition, that of no intercept point, can be recognized since the velocity of the target along the LOS (DM_{rt}) will be greater than the velocity of the torpedo along the LOS (DM_{ra}).

To recognize the second condition (intercept point out of range), the running time (T_a) must be calculated. The rate of closure is equal to the torpedo velocity along the LOS (DM_{ra}) minus the true target velocity along the LOS ($DM_{rt} = DR - DM_{to}$), i.e., range rate minus own ship velocity along LOS. The running time is equal to the closure rate divided by the range at the time of firing (iR).

$$T_a = \frac{DM_a \cos L - (DR - DM_{to})}{iR}$$

$$= \frac{DM_a \cos L - DR + DM_o \cos B}{DM_a}$$

Since $L = \sin^{-1} \frac{R \times DB - DM_o \cos B}{DM_a}$

$$\cos L = \sqrt{1 - \left[\frac{R \times DB - DM_o \cos B}{DM_a} \right]^2}$$

This leaves two equations to be solved:

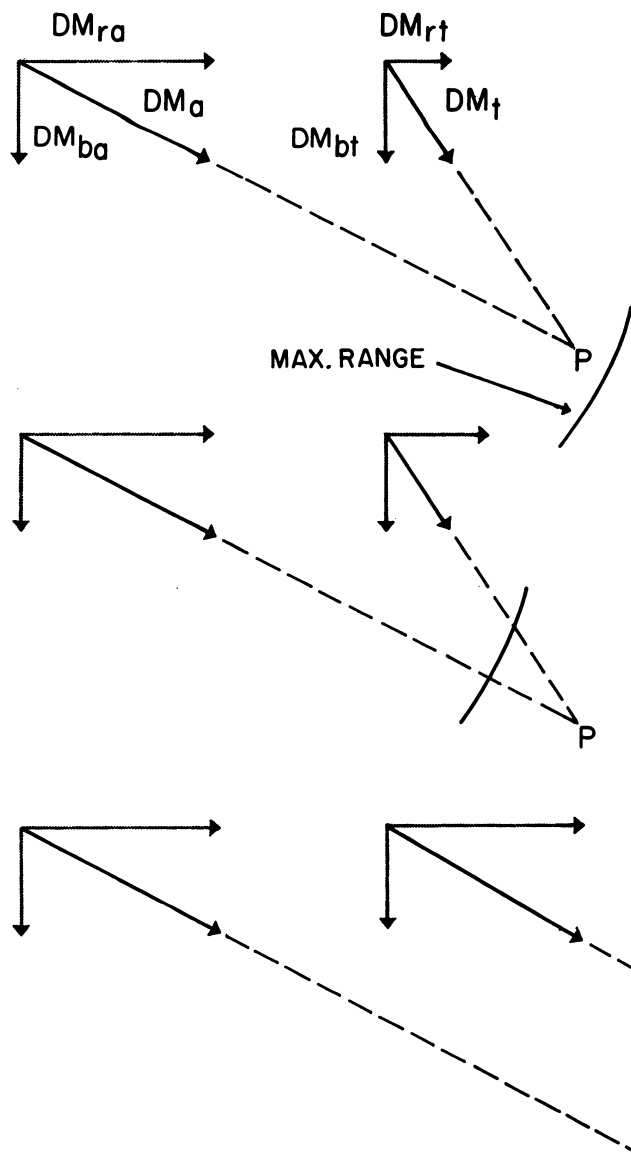
$$\text{Tube train Bdg}' = B + \sin^{-1} \left[\frac{R \times DB - DM_o \cos B}{DM_a} \right]$$

Running Time

$$T_a = \frac{DM_a \sqrt{1 - \left[\frac{R \times DB - DM_o \cos B}{DM_a} \right]^2} - DR + DM_o \cos B}{iR}$$

If the running time (T_a) is greater than the running time for maximum range ($T_{a \max}$), the intercept point is out of range. If T_a is negative, the target is moving away too rapidly to hit.

The values of R , B and DM_o must be made available to the computer in digital form. Since tracking devices are usually analog in nature, an analog-to-digital converter would be used to convert the measurements to digital form and provide this data as an input to the computer. The bearing rate $DB \text{ dB/dt}$ and range rate $DR \text{ dR/dt}$ are also necessary to the calculations. Differentiation is easily performed by a digital computer. Since $df(t)/dt = \lim_{\Delta t \rightarrow 0} \frac{f(t+\Delta t) - f(t)}{\Delta t}$ by definition; the computer takes one measurement (R_t), then after a time interval of, say 10^{-3} second, it takes another value ($R_{t+\Delta t}$) and $dr/dt = \frac{R_{t+\Delta t} - R_t}{10^{-3}}$



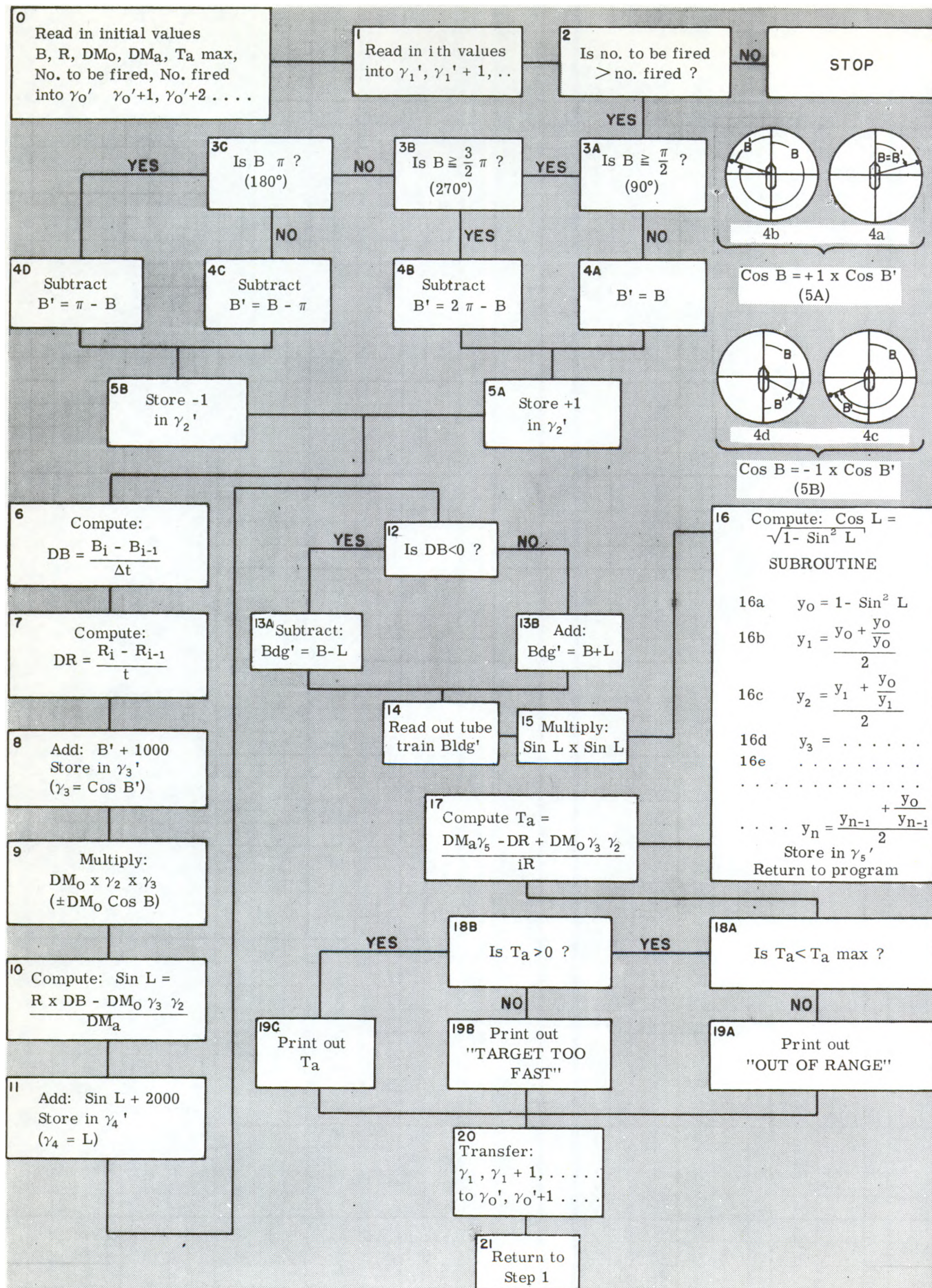
If the rate is constant there will be no error and if the rate is changing (target acceleration), the error can be made as small as desired by making the time interval (Δt) very small.

The solution also requires evaluation of cosine and arcsine functions. Converging series could be used for this process, e.g.

$$\cos x = 1 - \frac{x^2}{2} + \frac{x^4}{4} - \frac{x^6}{6} + \dots$$

$$\sin^{-1} x = x + \frac{x^3}{6} + \frac{1.3}{2.4} \cdot \frac{x^5}{5} + \frac{1.3 \cdot 5}{2.4 \cdot 6} \cdot \frac{x^7}{7} + \dots$$

but in a special purpose computer, where the same problem occurs many times, it is usually easier to provide a table. Memory locations 1000 through 1786 (.786 radian = 90°) would be reserved for a table of cosines. The cosine of .001 would be stored in location 1001, cosine .002 in location 1002, and cosine .786 in location 1786. The address of $\cos x$ is, therefore, $1000 + x$. A similar arrangement for computing arcsines would make the address of $\sin^{-1} x = 2000 + x$.



The problem also requires extraction of a square root. A computer can be programmed to extract a square root by the algorithm learned in high school. Iterative processes are preferable, however. A method of successive approximation is usually sought. This is a process whereby the computer makes a guess as a first approximation, uses this guess to get a second approximation, uses the second to get a third, etc., getting closer to the solution each time. One method for finding $y = \sqrt{x}$, where y_i is the i th approximation is:

$$y_0 = x$$

$$y_i = \frac{y_{i-1} + \frac{x}{y_{i-1}}}{2}$$

This converges very rapidly to \sqrt{x} . For example, if $x = 9$; $y_0 = 9$, $y_1 = 5$, $y_2 = 3.4$, $y_3 = 3.11$, $y_4 = 3.000$, $y_5 = 3.0000$.

The computer program can now be diagrammed. Two additional pieces of data are furnished so the computer will know when to stop, i.e., the number of torpedos to be fired and the number already fired.

The block diagram shows the procedure which will yield the required information. Steps involving purely arithmetic processes have been combined into one step to keep the diagram within reasonable limits. Step 6, for example, would actually require 1 subtraction and 1 division.

The "readout" of the tube train angle would normally result in the tube train angle being applied to a digital-to-analog converter. The analog output would then drive a servo system connected to the torpedo tubes. The "target too fast" and "out of range" information might be displayed in the form of a lighted indicator, and the running time converted to decimal notation, and displayed.

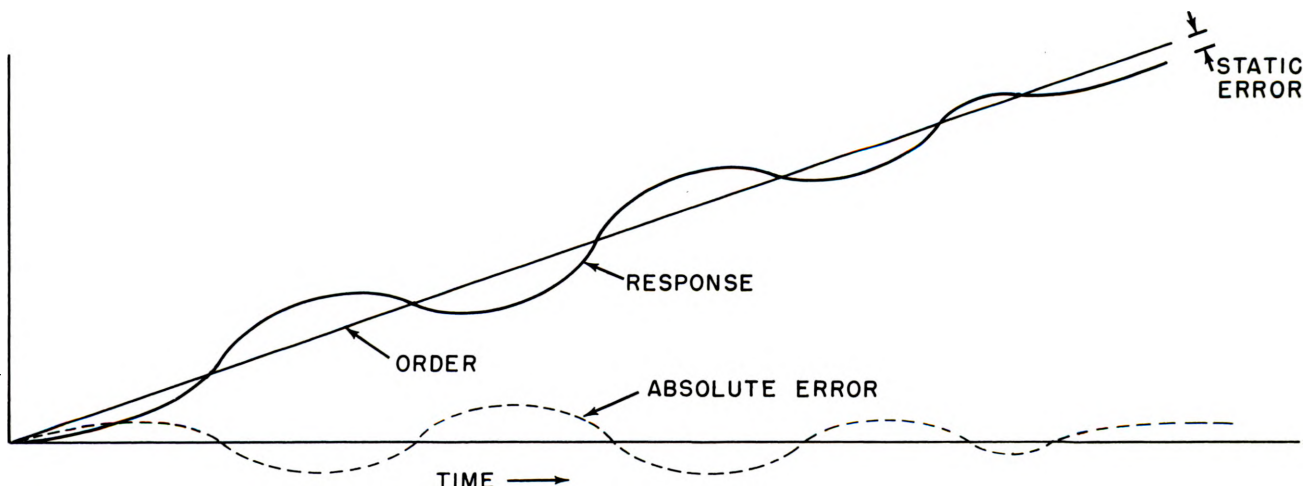
Contrasting the digital computer procedure for the solution of the tube train problem with the analog solution of the same problem (volume 1) reveals that the digital computer would be many times more complex than the analog computer. In fact, a digital computer would be very extravagant in this application. To understand which application is more suited to each type of computer, it is necessary to restate and compare the characteristics of analog and digital computers.

CONCLUSION

analog vs digital computers

Any comparison is complicated by the fact that the two techniques are fundamentally different, even though both may be used to achieve the same end result. For example, consider the apparently simple criterion of comparative accuracy. For a digital computer, accuracy is simply defined as the measure of the deviation of the output quantity from the exact value of the quantity. Accuracy in an analog computer depends on the use which is made of the output. The illustration shows the response of an analog computer to a ramp function. The actual computer output will respond in a similar manner,

oscillating about the exact value briefly, and then settling down to a value close to the exact value. If the computer is being used for engineering calculations, it is the static error, i.e., the error remaining after the output has stopped oscillating, which would be of concern. If the output were being used to position an object, such as a launcher or gun, the dynamic error would be of concern. For other applications, either the worst case error or the average error might be the significant quantities. Thus, a general comparison of analog and digital computers is not possible. Each factor must be considered in regard to a particular problem.



size

The most significant characteristic of digital computers, when weapons system applications are considered, is that there is a minimum size required for even the simplest computation. The memory for a simple digital computer can be small, but reading, writing, and access equipment is required, which is basically as complex as that for a very large memory. An arithmetic unit capable of at least addition and subtraction, and input/output devices, are essential to all digital computers. Therefore, a digital computer cannot be reduced beyond a certain minimum size, no matter how simple the problem it is required to handle. Analog computers can be made as small and simple as the problem will allow.

The size of an analog computer increases almost directly as the complexity of the problem, while the digital computer can handle larger problems with a small increase in size.

precision

The next most significant difference in analog and digital methods is the relative cost of different degrees of precision. (In this sense the "cost" of a given ability includes not only the price in dollars, but also any increase in solution time, in size, weight, susceptibility to adverse environment, failure rate, etc.)

In practice, the analog computer is usually a three-decimal (1%) machine. At considerable cost, four-decimal (0.1%) precision can be obtained when necessary. Beyond this point, the cost increases beyond all proportion to any increase in precision, and increased precision is often unobtainable at any price. In a digital computer, the cost of increased precision is approximately proportional to the amount of the increase. In on-line applications, where solution time is critical, precision exceeding the most demanding weapons system requirements is easily obtained. When solution time is not critical, the precision of a digital computer is unlimited, because the problem can be handled in parts.

solution time

The question of solution time is not as clear-cut as most other criteria of computers. The simplest digital computer capable of solving a given problem will usually require more time for the solution than the simplest analog computer, when the problem is in the field most often encountered in weapon systems, i.e., differential equations and trigonometry. On the other hand, a more complex digital computer is often faster than the analog computer. In any case, the question is often academic since even the slowest digital computer can usually be made fast enough. The comparison of solution time is further complicated because of the fundamental differences in the nature of the output data. The analog computer produces continuous output data. While the interval between the time the first input data is available and the time the first usable output data is available may be large, any changes in the inputs are very quickly reflected in the output data. The digital computer requires the same amount of time to process the first input data as it requires to process all subsequent inputs.

flexibility

Three distinct types of flexibility are desirable in weapons system computers. The ability to handle different problems in one computer, the ability to handle one problem in different ways, and the ability to expand to meet new requirements. In the first two cases the advantage lies with the digital computer. The ability to handle several problems at once is most easily accomplished by time-sharing, i.e., the computer performs a computation of the first problem, switches to the second, then to the third, etc., and finally back to the first. Such time-sharing is possible with analog computers, but seldom practical. Each variable in an analog computer is represented by some physical quantity, voltage, shaft position or such. When the inputs are switched, each quantity must change to correspond to the new data. This might require 180° rotation of potentiometer shafts or several hundred percent change in voltages. No such problem arises in a digital computer; the time required for a solution is not dependent on the change in value of the quantities involved. A second problem encountered is that it is difficult to recognize the completion of a problem in an analog computer.

A digital computer produces one result per cycle. When a result is obtained, the computer is finished with one set of input data and can be switched to another. The analog computer produces output data at all times; therefore, it is difficult to determine when to switch to another problem.

The second requirement for flexibility arises from the need to alter a computer to correspond to changes in weapons or tactics. Unless the change is a very minor one, such changes may require extensive alterations in an analog computer. Any changes in weapons can usually be accommodated in a digital computer by the simple operation of changing the data stored in the appropriate memory locations. Changes in tactics may be handled by changing the program. The third area in which flexibility is desirable is imposed by the possible need for expansion. With an analog computer, new components must be added, one at a time as required. While space and weight limitations may limit expansion, it is usually a fairly simple process. Once the capacity of a digital computer is reached, it cannot be increased without either major redesign of entire sections or a large increase in solution time.

reliability

The analog computer has proven more reliable than simple digital computers in the past. Conversely, digital computers with extensive self-checking and error-detecting features are usually more reliable than analog computers, but at the cost of considerable increased complexity. The superiority of analog computers with respect to reliability is not primarily due to the nature of analog devices, but to experience. Since analog devices have had a longer and more widespread development, more experience has been gained in producing reliable components. This is an advantage which can be expected to decrease in time, as development of digital devices continues.

capabilities

The area of capabilities is one in which few conclusions can be drawn. Almost all mathematical problems which can be handled by one type of computer can be handled by the other. The only generalization

which can be made is that the analog computer is better suited to handling single problems involving continuous quantitative computation, while the digital computer is better suited to handling multiplex problems and problems involving logical decisions.

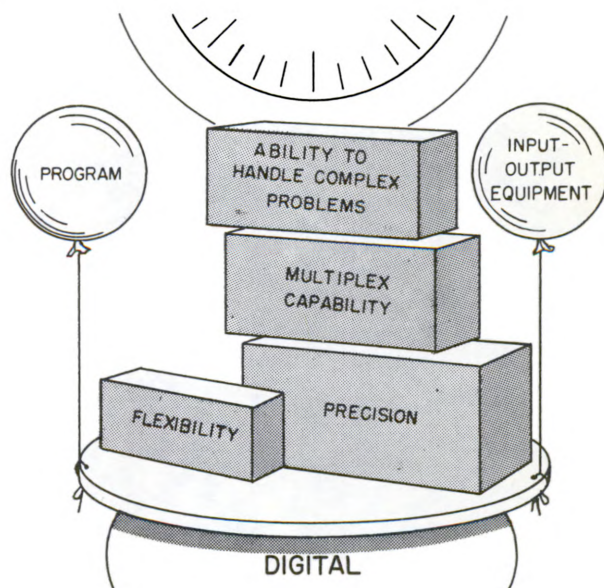
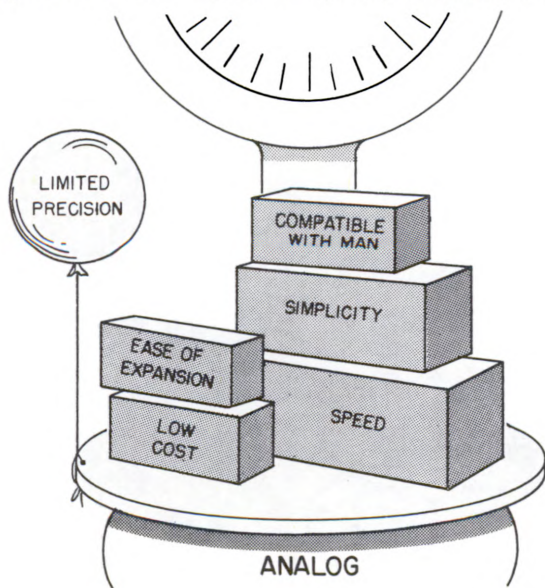
summary

The question of which type of computer is best for a given weapons system application is one that can seldom be answered without a detailed study of the requirements.

The only requirement which can immediately rule out one form of computer is the requirement for precision greater than that which can be obtained with an analog computer. Except in the area of guiding long-range missiles, and navigation, such precision is rarely required in weapons systems. The evaluation of computers is further complicated by the changing value of different characteristics. Cost, for example is a variable of constant value. A dollar saved is always worth precisely one dollar. Precision and solution time are variables which are valueless beyond a certain point. Once the precision of the computer equals the accuracy of the measurement involved, any further gain has no merit. Similarly, if the computer can solve a problem fast enough, any further gain in speed is not useful. Other characteristics are always worth something, but become less valuable as further gains are made. For example, the ability to expand is useful, but as this ability increases it becomes less and less likely that the capability will ever be utilized. These factors, plus the individual problems involved in each application, usually make it necessary to perform preliminary designs of both types of computers before deciding which is better for a particular system.

It can be said that highly complex problems and problems involving multiplex inputs are favorable applications of a digital computer, while small-scale

problems are favorable applications of an analog computer. Historically, analog computers have been predominant in naval weapons systems. Several factors have combined to favor the analog computer. The class of problems relegated to computers have been almost exclusively quantitative calculations of continuous variables, i.e., range, lead angle, etc., and the environment, usually shipboard or aircraft, is extremely unfavorable to sensitive digital components. The logical decisions have been reserved to man. As the increased speed and effectiveness of modern weapons quickens the pace of warfare, the unaided man is becoming too slow in making logical decisions in many situations. Computers capable of presenting rapid, accurate and comprehensive displays of tactical situations to enable men to make rapid, well-informed decisions, as well as computers capable of making decisions themselves, are becoming essential. The development of more accurate and longer range measurement techniques and weapons has made greater precision in computers more desirable. Both these developments have extended the area of application of digital computers in naval weapons systems. As more and more computers are employed, the possibility of using a single digital computer to handle, on a time-sharing basis, the calculations otherwise performed by several analog computers becomes more attractive. While it seems unlikely that digital computers will supplant analog computers in those equipments in which the analog is already used, many of the newer applications of computers will probably be digital.



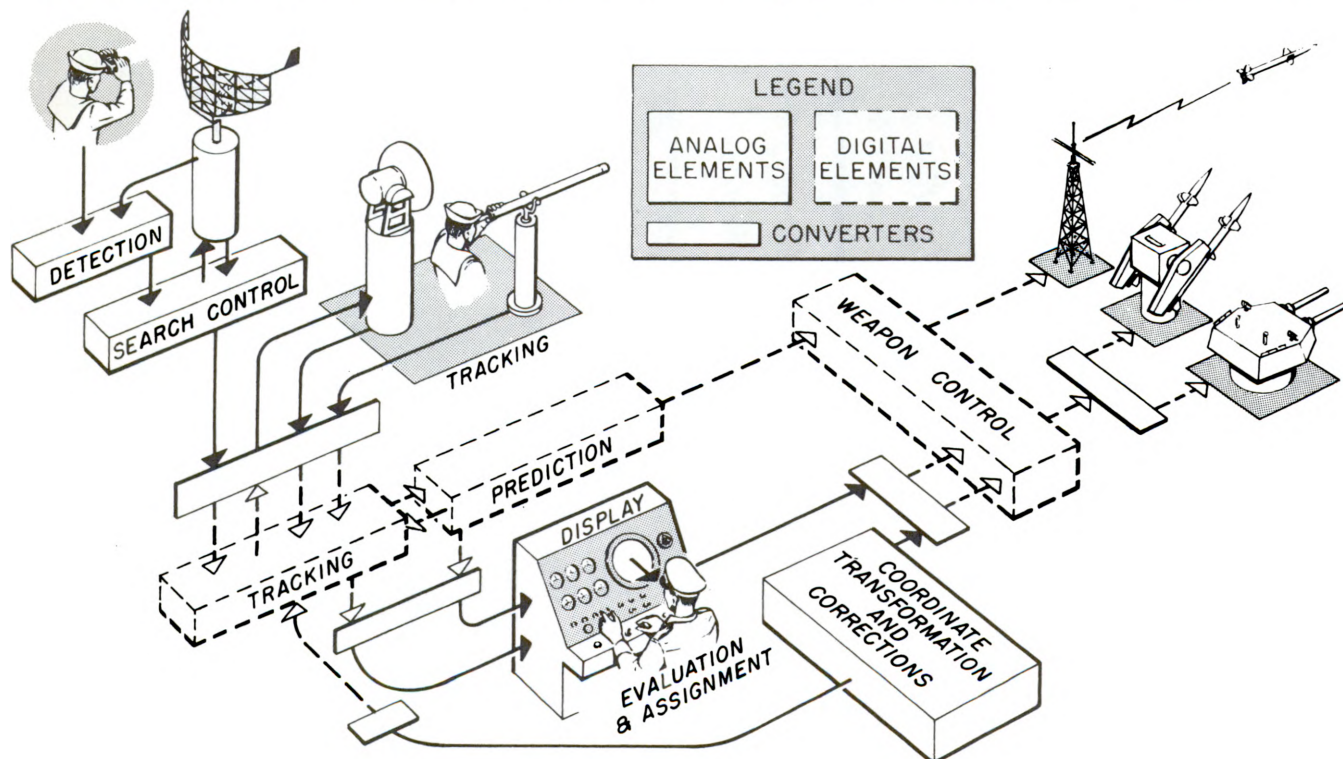
COMPUTER APPLICATIONS TO

Computers already perform many functions in naval weapons systems, particularly in fire control, and can be expected to perform many more functions in the future. Navigation, tracking, prediction, weapon control and guidance, and tactical display problems have been handled by computers. Other, less dramatic, applications include the handling of logistic problems, and automatic test equipment to check-out complex weapons system components. These computers occur in many forms. One very large computer may be used for many jobs; for example, tracking and predicting for many targets at once, or performing guidance computations for several weapons of different types at the same time. This offers an advantage if a digital computer is used, since the large computer can be much smaller than the total size of many lesser computers which would otherwise be required. If an analog computer is used, no such advantage is gained. The major disadvantage of a centralized computer remains, however: damage to the computer can render ineffective all the weapons controlled by that computer. Another disadvantage of the centralized computer is that any change in one weapon may require alteration of the entire computer. The first drawback can be reduced by using a parallel or standby computer which can take over in the case of a failure in the primary computer. Designing a computer as flexible as possible reduces the second drawback.

Since analog techniques are particularly suited for continuous quantitative calculations, and digital techniques for discrete logical calculations, the use of both types in the same system appears attractive. As pointed out earlier, however, the necessity for converting data from analog-to-digital and digital-to-analog form, which this entails, seriously limits

the advantages gained by the combinations. Nevertheless, the combined use of both techniques is practical for many systems, in some cases because conversions must be made anyway, and in other cases because the gains outweigh the losses. The input data for fire control systems are almost always in analog form. Tracking data must also be in analog form to be useful. Therefore, if a digital computer is used, data conversion is necessary whether an analog computer is used or not. It is also usually necessary to introduce a man into the system somewhere, and man is also essentially analog in nature, at least with respect to communication. For example, consider the problem of providing data on multiple targets. A computer could print out, in digital form, the essential data on each target, i.e., range, bearing, altitude, velocity, etc. However, it would be impossible for a man to evaluate this data for more than two or three targets within the time allotted and get a comprehensive idea of the situation. It is necessary that the data be converted to analog form and displayed in a pictorial manner to be effectively interpreted by a man.

A hybrid computing system as it might occur in a complex fire control situation is illustrated. The tracking, prediction, and weapon control (launcher bearing, guidance, etc.) calculations are performed by digital computers, while the search control, detection, display, coordinate transformation, and target evaluation and weapon assignment calculations are performed by analog computers or men. This combination makes use of the multiplex ability of digital computers, in order to handle many targets and many weapons simultaneously, while accommodating the analog nature of man. The essential trigonometric calculations of searching and coordinate transformation are handled by analog methods also. Note that the system requires two analog-to-digital and two digital-to-analog converters.



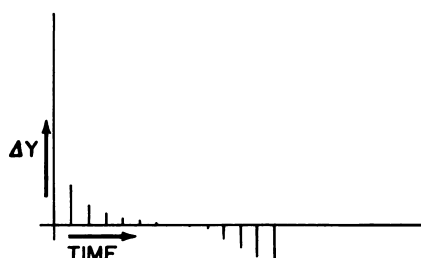
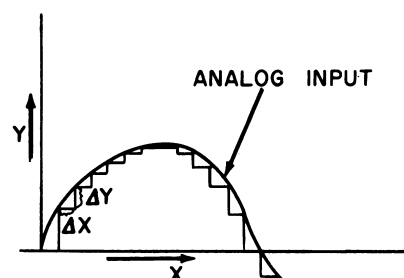
WEAPONS SYSTEMS

Analog and digital techniques may be combined in one computer as well as in one system. Computer components can be separated into two functional groups: the mechanisms which perform the actual mathematical operations, and those which set up the problem and control the operation. Hybrid computers may have 1) digital mathematical operators connected by analog techniques, 2) analog mathematical operators connected by digital techniques, or 3) both analog and digital components in either or both categories.

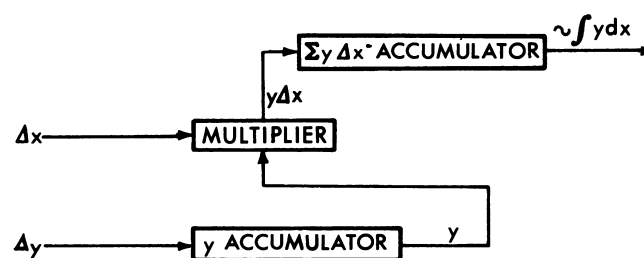
A specific example of the first combination, i.e., digital mathematical operators connected by analog techniques, is the digital differential analyzer (DDA). This serves the same purpose, and is used in the same manner, as the general-purpose analog computer. Instead of analog integrators, summing amplifiers etc., however, it uses digital components, digital data, and produces digital outputs.

This combination gains the precision of digital techniques, while retaining the most important advantage of analog computers in general purpose use, i.e., ease of programming, maintaining a one-to-one correspondence between elements of the computer and elements of the system under study, and allowing changes to be made while the problem is being run.

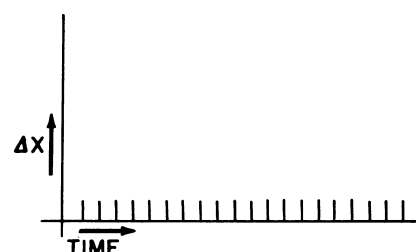
While the analog computer uses the magnitude of a function as the input data, a digital differential analyzer uses increments in a function as the input data. Given a function, $y = f(x)$, the input to an analog computer would be the value of y at all times. The input to a DDA would be a series of discrete signals, in digital form, each one representing the change in y since the previous signal (Δy). An in-



INCREMENTAL INPUTS



tegrator for a DDA, consisting of two accumulators and a multiplier, is illustrated. The problem starts with the appropriate initial value of the dependant variable (y_0) in the Y accumulator. Each succeeding value of Δy will be added to the previous total so this accumulator will retain the value of y at all times. As each value of Δx is received, the product $Y\Delta x$ will be formed and added to the total in the second accumulator. This accumulator, therefore, retains the value of $\Sigma y\Delta x$. Since $\Sigma y\Delta x \approx \int y dx$, the result is integration with respect to x . The variable $y\Delta x$ is also available to be used as the input to other integrators. Note that this type of integrator can integrate with respect to any variable, unlike the operational amplifier which can integrate only with respect to time. Note, also, that the results are stored (in the $y\Delta x$ accumulator), and can be read out at will. The complexity of the components of a DDA can be greatly reduced by taking increments of a unit value of one of the variables. If a new measurement of y is taken for each unit change in x , for example, the value of Δx will always be 1. This will enable the multiplier, in the integrator illustrated, to be eliminated, and an AND gate substituted. Each increment of x will cause the corresponding value of $y\Delta x$ to be added to the total in the $\Sigma y\Delta x$ accumulator. This technique transfers a characteristic advantage of analog devices to digital components.



One of the factors which tends to give analog computers short solution times is the fact that, with continuous inputs, the value of each variable at a given time will differ only slightly from the value at a previous instant. The analog computer takes advantage of this situation, and the digital computer does not. On a digital computer, a given calculation takes a fixed amount of time regardless of the previous calculation. Taking measurements at unit increments of x , in this example, eliminates a multiplier and increases the speed of the integrator by a large factor.

The digital differential analyzer is primarily intended for general engineering uses. The techniques, however, have many advantages which can make application to weapons systems desirable. A situation often encountered in weapons systems is the relatively simple problem requiring great precision. The digital computer has a characteristic minimum cost; which may be unreasonably high for a simple problem, but an analog computer with the required precision may be impossible to build. Using the techniques of the digital differential analyzer, the computer can be made as simple as is compatible with the problem. As a bonus feature, the time-sharing procedure of digital computers can also be applied.

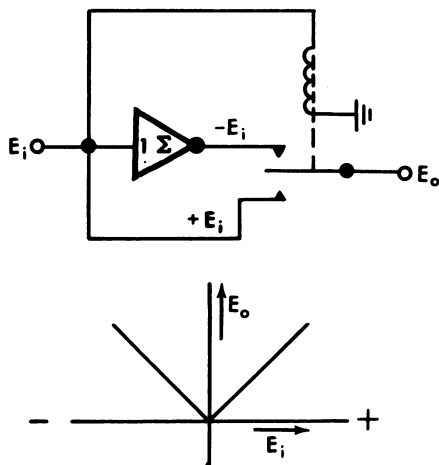
The second configuration, i.e., analog computing elements, connected by digital techniques, also has attractive advantages for weapons system applications. The problems occurring in weapons systems, particularly in fire-control, are inherently analog, and basically similar - so similar in fact, that the computer for one problem will often vary in only a few details from the computer for another problem. Unfortunately, the problems are not usually sufficiently alike to use exactly the same computer setup. By using digital switching to set up the analog components for whichever arrangement is needed each time a problem arises, the advantage of analog devices is retained without the disadvantage of characteristically long set-up time. The drawback lies chiefly in the fact that only one problem can be handled at any one

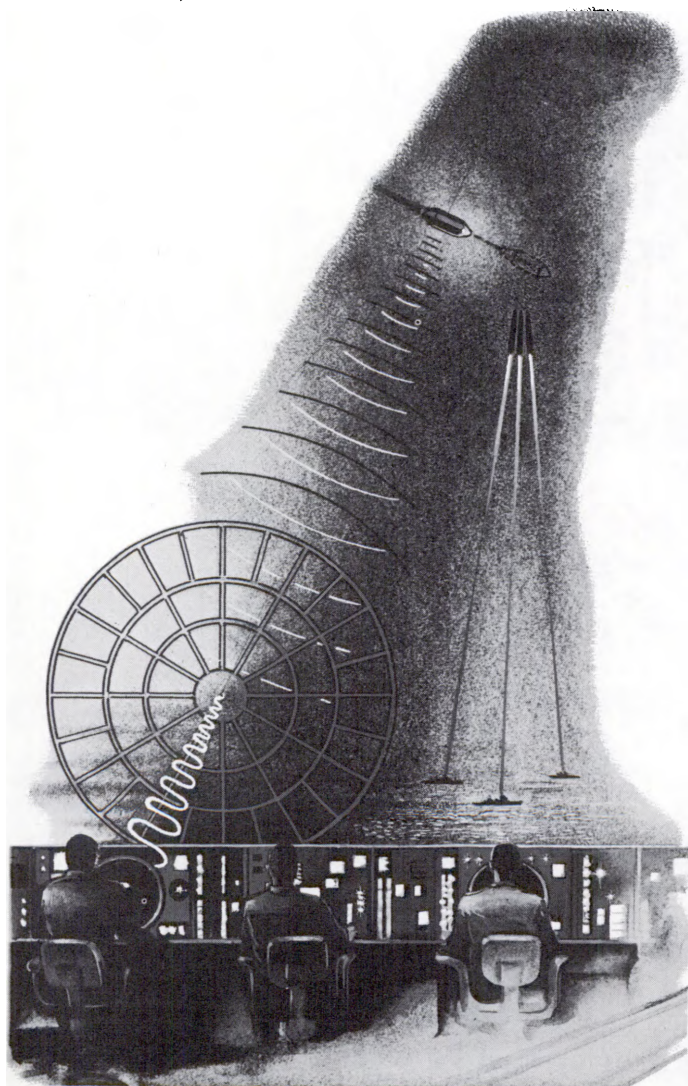
time. This may not be important in many cases. For example, a ship might have a single launching and guidance system capable of handling either of two types of missiles. The computer requirements for guiding each missile may be different, hence the components will be rearranged according to the missile in use, but only one type of missile is used at any one time, so only one computer is needed. Another application is guiding interceptor aircraft on a pass at a target. There are several different types of approach used, each offering advantages in different situations. Since only one approach is used at any given time, only one computer is required, although it must be capable of attaining more than one configuration.

There are also some other devices which combine both analog and digital characteristics. For example, an absolute value function generator may combine digital switching with an operational amplifier. The operational amplifier inverts the incoming signal. The relay will push out the contact when the input is positive, and pull in the contact when the input is negative. The output (E_o) will be equal to the input (E_i) when E_i is positive; and equal to $-E_i$ when E_i is negative.

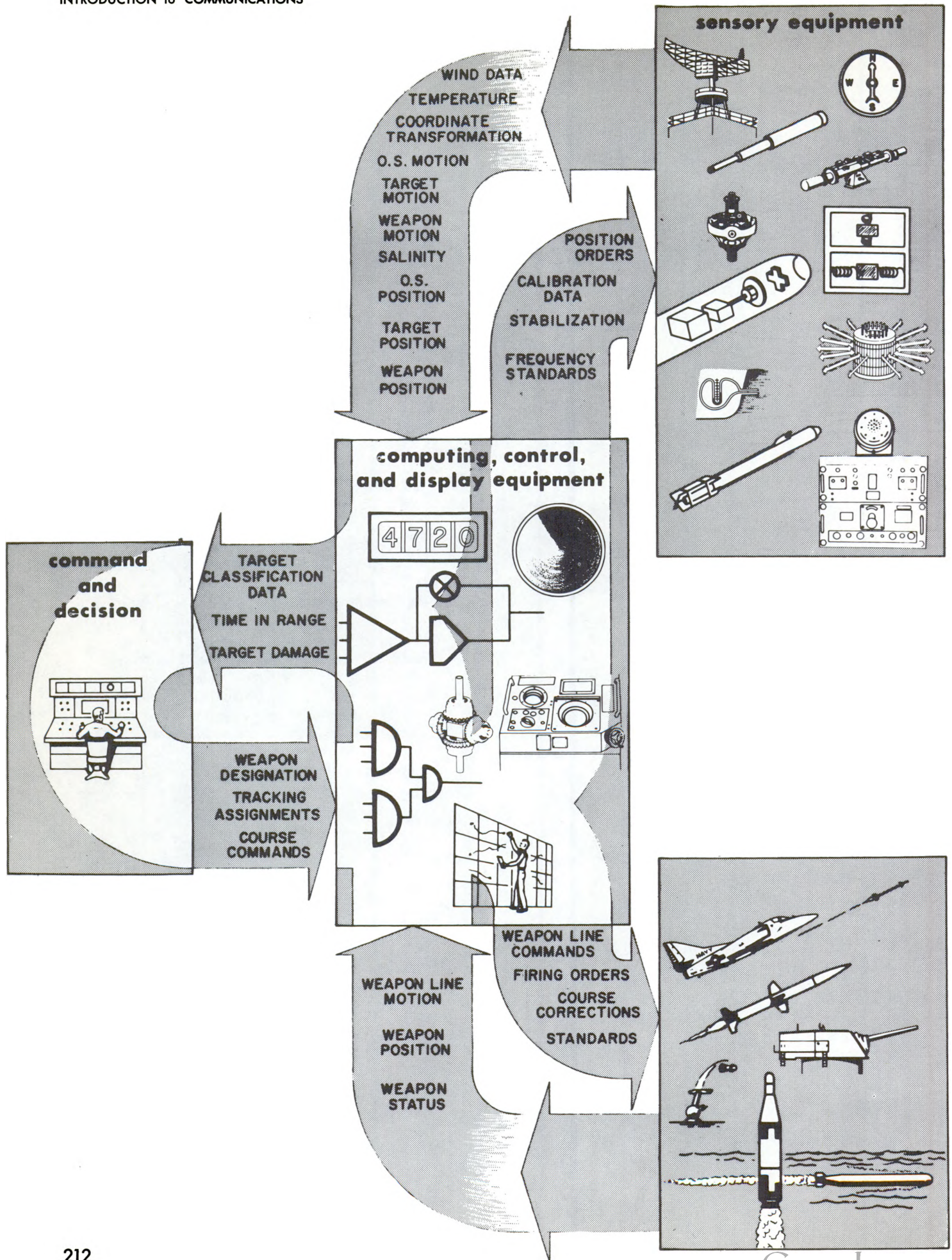
The application of computer techniques to automatic test equipment has the most clear-cut distinction between logical control and analog measurement. The method of testing any device is a logical, step-by-step procedure, as follows: Determine what the device is supposed to accomplish. Couple a simulator to the normal input points. Couple a measuring device to the normal output points. Simulate the normal inputs. Measure the outputs. Compare the actual output with the proper output. If the actual output is found to be correct, the device is operating properly; if it is found to be incorrect, proceed to isolate the fault by selecting a smaller section of the device, and return to the first step. When there are no smaller sections, the section under consideration is at fault.

This is a logical procedure ideally suited to digital techniques. The actual testing, however, requires simulation and measurement, both clearly analog (except in the case wherein digital inputs and outputs are normal, but this is a special case, and not only the digital sense, but also the actual magnitude and duration of such quantities would normally be measured in testing). The only quantitative operation of a digital nature involved is that of comparison. By using a hybrid computer to perform these operations, many desirable features of both analog and digital methods are combined. Since the programming is digital, sufficient versatility is gained so that many different devices may be tested with the same computer by simply inserting a different program. The ability to execute branch instructions is also valuable in this application, since each step will depend on the result of the previous step. A hybrid computer in this application can have speed and versatility which cannot be achieved by an analog machine, at a lower cost than that of a digital machine.



*Introduction to***COMMUNICATIONS**

A vital task in any system is the function of providing data transmission throughout the interrelated components of the system. Of extreme importance is the necessity for accurate and concise transmission of information between machines and between man and machine. The need for communication, the most efficient and economical methods of utilizing communications equipment, and the techniques available and how they are implemented are factors that lead to an optimum design specification for a communications system.



COMMUNICATIONS SYSTEM REQUIREMENTS

COMMUNICATIONS SYSTEM REQUIREMENTS Typical requirements for communications in a weapons system are: transmission of target location and classification data to decision levels and computing elements, transmission of tracking data to computing and control elements, delivery of coordinate transformation data to all areas requiring it and transmission of all sensory data, i.e., radar, sonar, optical, infrared, etc., to decision and computing elements.

The flow of information in a typical shipboard missile weapons system is illustrated.

The transmission of information requires communication between man and machine. Segments of the information are in analog form and part in digital form. Even when the form is the same, the language may be as different as the pulse data of a radar set and the graph from a bathythermograph slide. The rate may vary widely, from the very slow speed of positioning a ship to launch depth charges, to the very high speed necessary when supplying guidance data to a guided missile. Equipment such as inertial sensors may produce a steady flow of data, while other equipment such as a digital computer may produce nothing for long periods, and then yield a sudden burst of information.

The communications equipment must adapt to all these variations in order to make the system components compatible. This may require converting data from one form to another, or storing data until it can be accepted by a low speed component. Furthermore, the communications system must be able to handle multiple inputs and remain operative in spite of noise and possible countermeasures.

One might venture to say, "Why not design a system where both language and speed are the same for all components of the system?" This approach while feasible in small system, is highly impractical if not impossible in large systems. For example, a typical search radar may penetrate to 150 miles and a height of 90,000 feet. This is accomplished with relatively heavy and stationary type equipment, easily detected and positioned by an enemy. Since the information gathered by the search radar is of such prime importance, the possibility of enemy vehicles seeking to attack and destroy the station is great, requiring the sensed information to be processed and classified as quickly as possible. The time limitation demands the use of a high-speed computer to perform the necessary procedural processes. Man must receive the information in such a form and manner that decision-making can be almost instantaneous. The language difference between machines is obvious when it is seen that the search radar delivers both video signals and synchro signals to the computer, and the computer must transform these pulsed signals into a coded digital form in its normal processing of the data. There is therefore need for an encoder-decoder link to act as a translator to match the language of the two machines.

Last but not least, communication design must consider the requirements of the man-machine link. The machine must be designed to feed man the correct type of information and at a speed he can handle. In the weapon system outlined above, man determines to a great extent the hostility or friendliness of a target detected by the search radar. Once this determination is made, in the case of a hostile target, he must decide how dangerous it is, if it should be attacked, and so, how soon and by what weapon. Sufficient information must be provided to him in spite of the fact that he cannot handle more than a few pieces of information at one time.

CLASSIFICATION OF COMMUNICATIONS EQUIPMENT

Communications equipment may be classified as follows:

BY MESSAGE FORM The form by which the equipment is addressed, such as voice, image, or code.

BY MEDIUM OF TRANSMISSION The medium by which the equipment receives the message form, such as voice, wireless, sound, infrared, semaphore, light.

BY METHOD OF TRANSMISSION The method by which the message is sent over the medium to the equipment, such as continuous signal, discrete, or intermediate (a combination of both continuous and discrete).

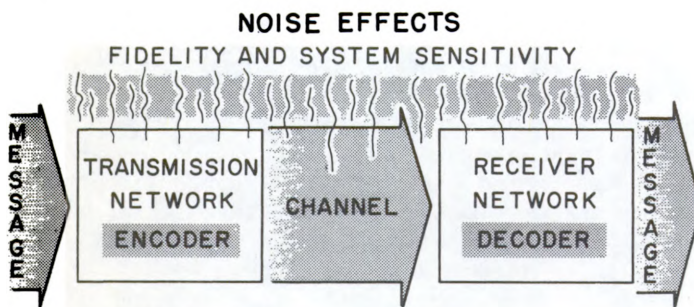
summary

The fundamental purpose of any communications system is to transmit information from one place to another. The nature of the system is determined by the type and quantity of information it must transmit, and its effectiveness is measured by the amount of information transmitted reliably.

A knowledge of basic information theory is essential in order to grasp the principles of communication.

INFORMATION THEORY

Information theory is a relatively new field. Its beginning took place in the mid-1920's when Nyquist and Hartley attempted to formulate theories on the capacity of a communications channel. In the late 1940's, a breakthrough in the field occurred when Shannon succeeded in formulating theories concerning the channel capacity and information content of a message. Shannon's theory ignored the meaning or value of a message and concentrated on the classes of messages as a group. Thus it was possible to reach definite mathematical conclusions which permitted the derivation of explicit formulae relating to such fundamental parameters as signal-to-noise ratio and bandwidth. This theory solved the question of the rate of transmission of information under specified circumstances. These circumstances revolved about the nature of the signal source, discrete or continuous; the nature of the channel, in particular its capacity for transmission; the nature of the noise perturbing the transmission; and the fidelity of the transmission.



The process of transmitting a message is illustrated by the schematic of a general communications system. The signal source provides the message which is to be transmitted. This message is then encoded, that is, a converter designated as the encoder translates the message from one language form to another, e.g., words or numerals are translated into a code. The coded information is then transmitted through a channel as a signal to the decoder. The channel could be a transmission line or any one of a number of carriers. A noise source is usually associated with the channel. Also there is a noise source associated with the encoder and the decoder. The total system noise can be represented as a single noise source as illustrated.

The decoder receives the signal from the channel and translates the code back into the original language or some other language depending on the destination of the received message.

The first requirement for a quantitative study of communications is a mathematical expression for the amount of information (I) contained in a message. There is no strict derivation possible since there is no definition of

"information content." The problem, therefore, is to find a mathematical function which fits our intuitive ideas of information content. Before anything else can be done, it is necessary to determine the variable of which the information content is a function:

$$I = f(?) \quad (1)$$

Consider two communications situations:

1. A commander is asked whether or not his ship has more than 24 missiles of a certain type on hand (assume a full complement is 49 missiles). Two replies are possible: "yes" or "no", and each is equally probable.
2. A Commander is asked how many missiles he has on hand. In the case there are 50 possible replies (0 through 49, inclusive).

Clearly the reply would contain more information in the second situation than in the first, since it would tell the recipient exactly how many missiles are on hand. This suggests that the information content of a particular message can be expressed as a function of the number of possible messages (n):

$$I = f(n). \quad (2)$$

The next step is to determine which function. Consider a message of two parts (1 and 2), each part having n_1 and n_2 possible values, respectively. Since the total message may have any combination of the components, there are $n_1 \times n_2$ possible messages. The total message must contain an amount of information equal to the sum of the parts:

$$I_{\text{total}} = I_1 + I_2, \quad (3)$$

$$\text{or } f(n_1 \times n_2) = f(n_1) + f(n_2). \quad (4)$$

The logarithm function has the required form:

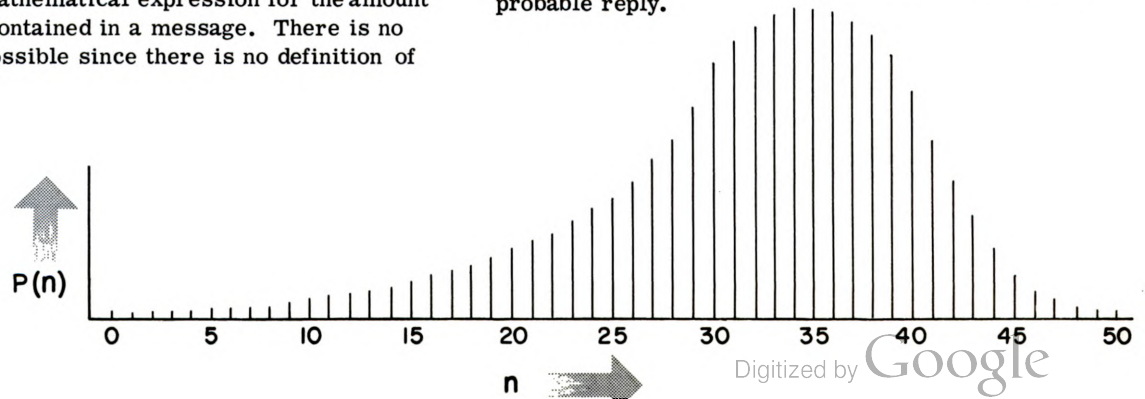
$$\log(x + y) = \log x + \log y. \quad (5)$$

Therefore, the expression:

$$I = k \log n, \quad (6)$$

where n is the total number of possible messages and k is a constant depending on the units, seems to fit the intuitive concept of information content.

However, a little further study will show that the information content is not the same for all messages. Returning to the communications problem — if a commander is asked how many missiles he has on hand, and the interrogator has estimated on a sound basis that the ship should have about 35 missiles on hand, then a reply of "35" is most probable, "34" or "36" is somewhat less probable, etc. If the reply is "35", it simply confirms what was expected. If an improbable reply such as "0" is received, it conveys more information than a more probable reply.



This suggests that the information content of a message in which all possible messages are not equally probable, can be expressed as a function of the probability of the message (P_i):

$$I = k \log \frac{1}{P_i} = -k \log P_i. \quad (7)$$

Note that this reduces to equation (6) in cases where all messages are equally probable, since $P_i = 1/n$ in such a situation.

Therefore:

$$I = -k \log P_i = -k \log \frac{1}{n} = +k \log n. \quad (8)$$

In this study, the information content of a particular message is of little interest. What is most important is the average amount of information of all messages (H). If the probability of message i is P_i , then this message should occur nP_i times in a series of n messages. Therefore, the average amount of information is given by:

$$H = \frac{+nP_1 I_1 + nP_2 I_2 + nP_3 I_3 \dots nP_n I_n}{n}$$

$$H = \frac{-nP_1 k \log P_1 - nP_2 k \log P_2 - nP_3 \dots - nP_n k \log P_n}{n}$$

$$H = - \sum_{i=1}^n P_i k \log P_i \quad (9)$$

This quantity (H) is frequently referred to as the entropy of information and is based on its analogy to thermodynamic entropy, which is a measure of degree of randomness, disorganization, or uncertainty.

Returning to equation (6), it is necessary to evaluate the constant k .

Since $k \log x = \log_k x$, equation (6) can be written as:

$$I = \log_k n. \quad (10)$$

The basic unit of information is 1 bit; defined as the amount of information required to distinguish between two equally probable events. This information can be contained in a message consisting of one of two possible symbols (1 binary bit).

Therefore:

$$1 \text{ bit} = \log_k 2; \quad (11)$$

and since $\log_2 2 = 1$, $k = 2$.

If a decimal number system is being followed, it is more convenient to use a unit of information equal to that necessary to distinguish 1 of 10 possible events. In this case:

$$1 \text{ unit} = \log_k 10, k = 10. \quad (12)$$

Two communication situations illustrate the use of entropy. A message consisting of two binary bits is transmitted, and each message is equally probable. Since there are four possible messages (00, 01, 11, 10), and each is equally probable, $n = 4$ and $P_1 = P_2 = P_3 = P_4 = 1/4$.

Therefore:

$$H = - \sum_{i=1}^4 P_i \log_2 P_i \quad (13)$$

$$= - \left[\frac{1}{4} \log_2 1/4 + \log_2 1/4 + \frac{1}{4} \log_2 1/4 + \frac{1}{4} \log_2 1/4 \right]$$

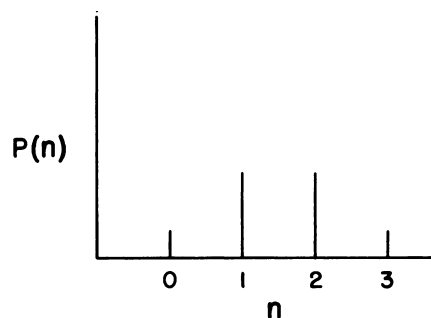
$$= - \log_2 1/4$$

$$= \log_2 4$$

$$= 2 \text{ bits per message.}$$

Since all messages have equal probability, this is the maximum entropy case. In general for n symbols H is greatest when all probabilities are equal to $1/n$.

A case in which all messages do not have the same probability of accuracy might arise if one were asked to flip a coin three times and transmit in binary code the number of heads that occur. This type of event follows the binomial distribution, 1, 3, 3, 1. Thus, if P_i is the probability of i heads:



$$P_0 = 1/8, P_1 = 3/8, P_2 = 3/8, P_3 = 1/8$$

$$H = - \sum_{i=0}^3 P_i \log_2 P_i$$

$$= - \left[\frac{1}{8} \log_2 1/8 + \frac{3}{8} \log_2 3/8 + \frac{3}{8} \log_2 3/8 + \frac{1}{8} \log_2 1/8 \right]$$

$$= + \frac{2}{8} \log_2 8 + \frac{6}{8} \log_2 \frac{8}{3}$$

$$= \frac{6}{8} + \frac{6}{8} \times 1.41$$

$$= 1.81 \text{ bits per message}$$

Aside from the amount of information and entropy of information, several other parameters must be formulated. These parameters are relative entropy, redundancy, channel capacity, and coding. However, before developing these parameters, a discussion of messages and their sources is necessary.

First, a source is defined as that which generates all possible symbols contained in a finite number of alphabet symbols. For example, the simplest possible source is one that generates symbols from an alphabet of two symbols; that is, if the alphabet symbols are 0 and 1, then the possible symbols are 00, 11, 01, and 10. From this simplest source, we can advance to more complicated sources merely by increasing the alphabet, by allowing symbols having different probability, by having successive symbols dependent, and finally by having symbols which require different lengths of time. When an alphabet symbol is transmitted along a communications channel, it becomes a message symbol. The repeated occurrence of message symbols yields messages. In the two-symbol alphabet above, there are four possibilities, 00, 01, 10, and 11, each of which has equal probability. In this case the entropy is maximum ($= H_{\max}$). From equation (15), we have seen that maximum entropy in an n symbol alphabet is $\log_2 n$, when the probabilities are equal to $1/n$. Whenever the entropy is not maximum (the source has an entropy H less than H_{\max}), then the ratio of H/H_{\max} is the relative entropy or the efficiency of the system. These two parameters always lie in the range 0 to 1 and greatly aid the information theorist to increase the efficiency of a code.

channel capacity

The capacity of a channel may be defined as the maximum rate a source or sources may generate symbols (symbols per seconds) and have them transmitted over the channel, yielding a maximum rate (in bits per second) at which information flows. This rate may be defined as information bits per unit time, and may be measured by the formula, $C = \frac{I}{t}$, where:

- I = is the average amount of information per message from the source
- C = channel capacity
- t = time in seconds

rate

It is now possible to state a basic theorem relative to any communications channel. When a source has average information content I (bits per symbol) and a channel has capacity C (bits per second), then the maximum rate of transmittal without error can never be greater than $\frac{C}{I}$.

As an example, consider four possible messages, A, B, C, and D, with probabilities of occurrence given respectively, by $P_A = 1/2$, $P_B = 1/4$, $P_C = 1/8$ and $P_D = 1/8$. The average amount of information per message, when evaluated using equation (13), is found to be: $H = 1/2 \log_2 1/2 + 1/4 \log_2 1/4 + 1/8 \log_2 1/8 + 1/8 \log_2 1/8 = 1.3/4$ bits per message.

Thus, on the average there are 1.3/4 bits of information per message transmitted.

Assume the communication channel includes an operator, at a switchboard composed of four switches that can be opened or closed by actuating companion keys labelled A', B', C', and D'. If the operator could activate one switch per second, it would then be possible to send one symbol per second. It has already been shown that the amount of information associated with a set of four messages is a maximum when all four messages have equal probabilities. Therefore, the maximum amount of information associated with each message of the optimum set having equal probabilities would be $H = -\log_2 1/4 = 2$ bits per message. As there are two bits of information associated with each symbol, the channel

capacity ($C = \frac{I}{t}$) is limited by the operator's speed of transmission. Accordingly, it is possible to encode the previous messages, A, B, C, and D into such a form that it would be possible to transmit these messages at a rate of $R = \frac{C}{I} = \frac{2}{1.3/4} = \frac{8}{7}$ messages per second, or 1.1/7 symbols per second.

coding

If there can be devised a method for increasing the rate without sacrificing reliability, then the effectiveness and efficiency of the communications can be increased. This can be accomplished by proper encoding of the messages to be transmitted. In coding, the idea is essentially to produce a set of symbols to be transmitted over the channel so that they will all occur independently and with equal frequency. If this has been done, then the output of the encoder will look like a source of maximum entropy, and the channel capacity can be fully utilized, if the source generates symbols at the

proper rate. Consider the previous example of the four symbols, A, B, C, and D, where the entropy of the source is 7/4 bits per message. The maximum entropy for this four symbol source is two bits per message, and relative entropy $H/H_{\max} = \frac{7/4}{2} = \frac{7}{8}$. Therefore redundancy

$1 - \frac{H}{H_{\max}} = \frac{1}{8}$. Consider coding these symbols for transmission in binary form. Then 00 = A, 01 = B, 10 = C, and 11 = D, requiring two bits per source symbol. However, we know that the maximum rate of transmission is 1.1/7 symbols per second; thus the entropy is 7/4 bit symbols per second. Using a code of 0 for A, 10 for B, 110 for C, and 111 for D, the number of bits utilized in transmitting S symbols then is:

$$\frac{S}{2} + \frac{2S}{4} + \frac{3S}{4} = \frac{7}{4} S$$

or 7/4 bits per symbol, and since this is equal to the entropy of our source, it is the most efficient code possible.

Redundancy should not be considered undesirable. It is true that the maximum amount of error-free information per unit time can be transmitted through a perfect noiseless channel only by a system of message symbols (code) having no redundancy. This is only of theoretical interest, however, since there are no perfect noiseless channels available. When the code is redundant, the correct message may be deciphered even though some message symbols are lost or altered: "Fr xmpl, ths sntnc cn b rd wth lttl dfclty lthgh th vwls (whch rprsnt 35 prcnt f th mssg symbols) hv bn mmttd." This is due to the redundancy of the language. The redundancy is extensive since the symbols vary greatly in probability of occurrence (the probability of an "e", for example, is approximately 0.13, while the probability of a "w" is approximately 0.02). The symbols are not independent since the probability of a certain symbol is affected by the symbols which precede it, e.g., the probability of a "u" following a "q" is 1, and the probability of an "h" following a "t" is 0.37. The number of errors which can be incurred without affecting the proper reception of the message is proportional to the amount of redundancy in the code. It is also effected by the extent to which the redundancy is organized to permit efficient error detection and correction. The use of error detecting and error correcting codes permits the use of redundancy to overcome errors in transmission. The simplest and most common is the parity check.

PARITY CHECK The parity check is usually employed when data is transmitted in binary form, although the principles apply to other systems as well. If a message normally consists of five 5-bit words in binary form and 25 bits are transmitted, the code presumably has no redundancy. By employing 30 bits to transmit the 25 bits of information, the code is made 17 percent redundant, but errors can be detected. The parity bit for each row is chosen so that the row including the parity bit will have an even number of ones. (It is also possible to use an odd parity check, of course, but even parity is conventional.) If any one bit is altered in transmission, the row containing that bit will be of odd parity, and the error can be detected. This type of check does not enable the recipient to determine the correct message, but any single error will be detected. Two errors in the

same word, or any even number of errors, will result in even parity, and the errors will go undetected. The probability of two errors in one word, however, is usually very small if the words are short. More complex codes enable an error to be corrected as well as detected so that the message may be properly deciphered.

0	1	1	0	1
0	0	0	1	1
0	1	1	1	1
1	0	0	0	0
1	0	0	1	1

DATA

0	1	1	0	1	1
0	0	0	1	1	0
0	1	1	1	1	0
1	0	0	0	0	1
1	0	0	1	1	1

DATA AND
PARITY BITS

0	1	1	0	1	1
0	0	0	1	1	0
0	1	1	0	1	0
1	0	0	0	0	1
1	0	0	1	1	1

DATA RECEIVED

ERROR

ERROR CORRECTING CODES

By adding parity bits to check each column as well as each row, an error can not only be detected but also isolated to a single bit. By inverting the erroneous symbol, the correct message is restored. In the illustration shown, an error is indicated in the fourth row and in the third column. The erroneous bit, therefore, is at the intersection. Changing the zero in this position to a one restores the original message. Also this code will detect, but not correct, any pair of errors. In fact, a minimum of four errors occurring in a rectangular pattern is necessary to escape detection. This code offers good protection, but the amount of redundancy is extensive.

Similar results can be obtained by using only five parity bits to protect a 25 bit message. The illustration shows a block of 30 bits numbered for convenience. Positions 1 through 25 make up the message, and positions 26 through 30 are the parity bits. Instead of a function of only five bits, each parity bit is a function of 14 bits.

0	1	1	0	1	1
0	0	0	1	1	0
0	1	1	1	1	0
1	0	0	0	0	1
1	0	0	1	1	1
0	0	0	1	0	

DATA AND
DOUBLE PARITY BITS

0	1	1	0	1	1
0	0	0	1	1	0
0	1	1	0	1	0
1	0	0	0	0	1
1	0	0	1	1	1
0	0	0	1	0	

DATA
RECEIVED

ERROR

ERROR

0	0	1	1	1	1
0	0	0	1	1	0
0	0	1	0	1	0
1	0	0	0	0	1
1	0	0	1	1	1
0	0	0	1	0	

NON-DETECTABLE
ERROR

This is tabulated below, where a 1 in the column corresponding to each parity bit shows which data bits are protected by that parity bit. Each bit is protected by a unique combination of at least two parity bits. An error in position 10, for example, will effect parity bits 27, 28, and 29. The coded message illustrated contains one error. The reader is invited to locate it. The degree of protection offered by this code is not as great as that offered by the preceding illustration, but is more economical. Such codes are widely used to detect errors, but are not generally used to correct errors, since it is usually easier to request a repetition of any erroneous message than to make corrections.

1	6	11	16	21	26
2	7	12	17	22	27
3	8	13	18	23	28
4	9	14	19	24	29
5	10	15	20	25	30

position
numbers

parity key

message
with
error

1	6	11	16	21	26
2	7	12	17	22	27
3	8	13	18	23	28
4	9	14	19	24	29
5	10	15	20	25	30

PARITY
BITS

DATA BITS

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
26	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
27	0	0	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	1	1	1	1	1
28	0	1	1	1	0	0	0	1	1	1	1	0	0	0	1	1	1	1	0	0	0	0	1	1	1
29	1	1	1	1	0	1	1	0	0	1	1	0	1	1	0	0	1	1	0	0	1	1	0	0	1
30	1	1	0	1	1	0	1	0	1	0	1	1	0	1	0	1	0	1	0	1	0	1	0	1	0

channel types

There are three types of physical channels; first, the discrete noiseless channel; that is, a channel with no associated noise and which transmits some sort of discrete messages, for instance, dots and dashes; the second type is the discrete channel with noise; and the third type is the continuous channel with noise. The two noise channels are the most common of the three types.

DISCRETE NOISLESS CASE. The fundamental theorem for the discrete noiseless case states that it is possible to encode messages for transmission over a channel at a rate R , the maximum value of which is C/I , where C is the channel capacity and I is the average amount of information per message from a source. The theorem further states that it is impossible to encode the messages so that they may be transmitted without error at a greater rate.

DISCRETE CASE WITH NOISE. The fundamental theorem for the discrete case with noise states that if the rate of transmission is less than the channel capacity, it is possible to encode a message for transmission over a noisy channel so that an arbitrarily small percentage of errors may be obtained at the receiver. The capacity of a noisy channel is defined as the value of the product of the rate at which messages are being transmitted and the amount of information transmitted over a noisy channel, when the transmitted message set is selected so as to maximize the received information.

CONTINUOUS CASE. In the continuous case, where the output of the source is an analog quantity such as voltage, current, displacement etc., it is a common procedure to attempt a treatment in a manner analogous to the handling of the discrete case. This is effected by quantizing the analogous quantity. Quantizing is accomplished by sampling the signal at discrete time intervals. This procedure is based on two facts, the first being that nearly all signals to be transmitted are band-limited, i.e., they contain no components outside a given band of frequencies. For example, it is known that in order to transmit intelligible speech, a band of frequencies approximately 3,500-cycles-per-second wide is needed. The second fact is the sampling theorem. This theorem states that for a continuous signal of form $G(t)$, which has no frequency component higher than W cps, function $[G(t)]$ can be completely determined by observing the value of the function at time (t) $0, \pm 1/2W, \pm 2/2W, \dots, \pm n/2W$. Therefore, it is necessary only to sample the continuous signal every $1/2W$ seconds.

However, the definition of entropy differs for the continuous case and the discrete case. Discrete sources were described by specifying the probabilities associated with each possible symbol. Continuous sources are described by specifying the probability-density function associated with the possible amplitudes. The probability-density function is a measure of the amplitude of a signal, which lies between amplitude values x and $x + dx$ where dx is a small change in x . Equation (9) shows the entropy of the discrete source as:

$$H = -\sum_{i=1}^n P_i \log P_i$$

In the continuous case, the probability-density function is expressed as $p(x)dx$.

In the discrete case, the entropy of a source was maximum when all the symbols had an equal probability of occurrence. This is not the case with the continuous source. Simply stated, a signal with amplitudes normally distributed has more entropy than any other continuous signal of the same average A.C. power. A signal which is band-limited in frequency and normally independently distributed, is called band-limited white noise. It is white because over its frequency band the probability-density function which determines its amplitude is constant in the frequency domain. The channel capacity of the continuous source is a function of the bandwidth, the average signal power, and the noise power. The mathematical derivation is quite complicated and depends on the particular case under examination. It suffices here to state that the channel capacity for a continuous source containing band-limited white noise (of power N) and intelligence signals of maximum average power S is:

$$C = W \log_2 \left(1 + \frac{S}{N} \right) \text{ bits per second. (17)}$$

This is a most important relationship in information theory and is based on the following assumptions:

- Signal source is limited by its average power.
- Noise is band-limited, white, and independent of signal.
- Channel is limited to bandwidth W .

The fundamental theorems indicate that a code can be derived allowing any source to produce information and send it over a continuous channel with a rate equal to the channel capacity and with arbitrarily small error. Increase in channel capacity can be accomplished by increasing signal power, and bandwidth can be exchanged for signal-to-noise ratio in a channel. By exchanging bandwidth for signal-to-noise ratio in a communication channel, it is possible to transmit more information by increasing W appropriately if enough signal power is available to keep the ratio of S/N from being too small. And on the other hand, if the allowable band of transmission W is too small, increasing the signal power sufficiently can maintain the same rate of transmission.

communication systems factors

Some measure of effectiveness of communications systems are: speed, capacity, accuracy, and cost. All of these factors are dependent on each other. Speed, or time delay; capacity, or number of messages which can be handled; and accuracy, or error rate, can be improved by increasing a) bandwidth, b) power, and c) noise reduction. These relationships are based on the fundamental theorem expressed in equation (17). The cost of increasing time, capacity, and error rate of a communications system can become exorbitant, and it is often necessary to choose between system performance and financial outlay. The reliability of a system is a measure of the probability of successful transmission and reception of data. The simplest design usually is the most reliable. Action against enemy countermeasures, however, often leads to redundancy, lessening of reliability, and increase in cost.

TRANSMISSION METHODS

digitalized codes for transmission

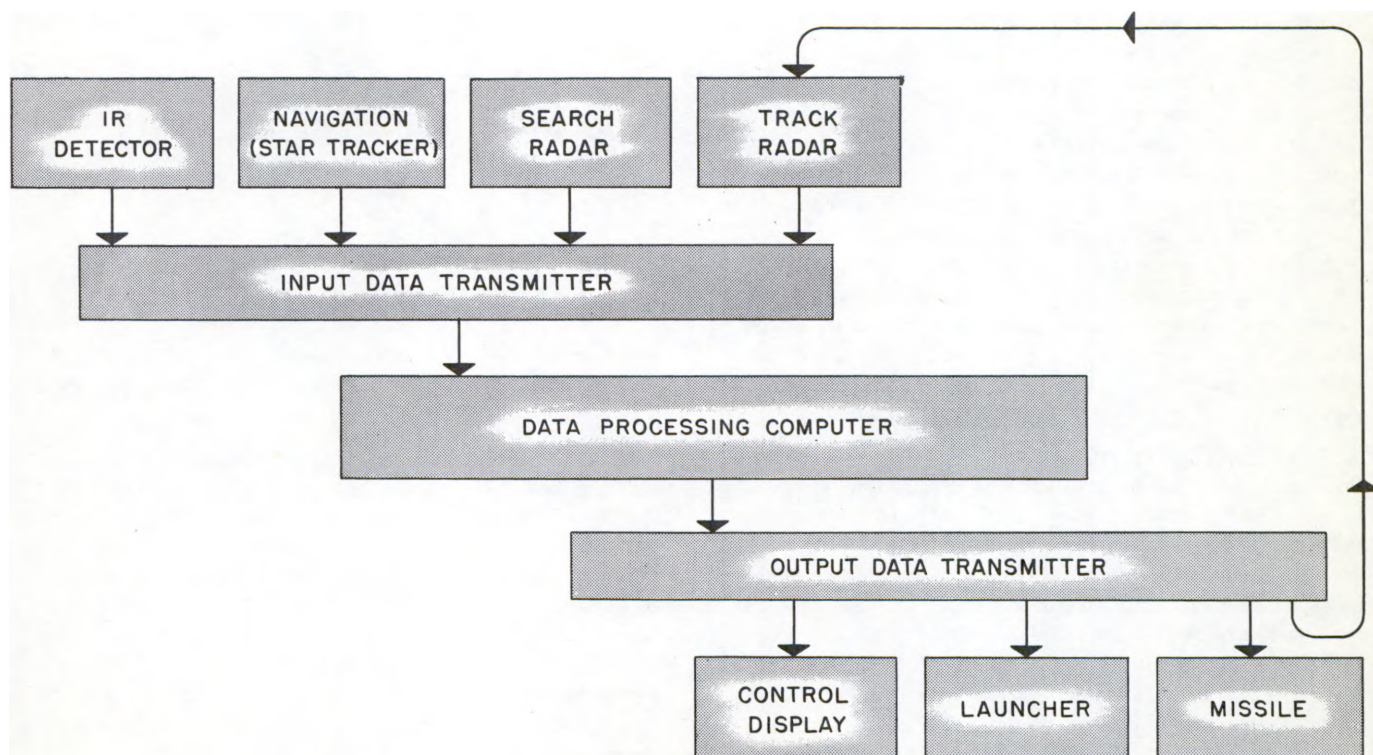
Consider a missile in flight which must continually receive information concerning both its own course and the target. If the target is not fixed and capable of evasive maneuvers, then the missile must be supplied with information concerning these maneuvers. Deviation from the previously determined positions must be quickly sensed so that guidance and control devices can be actuated before extreme measures are required. Analog information concerning changes in target course can be fed into a computer in coded form. The computer generates course commands in the same coded form. These commands can be transmitted in code to the missile where the information is decoded and utilized by the missile control system.

transits in short bursts

In this technique, intelligence is compacted and transmitted as very high speed bursts. This technique requires speed-up equipment for transmission and slow-down equipment at the point of reception. The method allows for higher security, capacity, and speed. Many times the reception of information bits at the transmitter requires a temporary storage of the bits for a short period of time until the buffer accumulates enough data to transmit a complete message. This information would then be transmitted in a short burst at high speed. Conversely at the reception end, the receiving buffer would reverse the procedure, store the bits, and only allow the flow of data to the indicator when a comprehensible message is realized.

converters

Converters are utilized to provide a conversion between two different types of analog signals or between an analog and digital signal. In many systems, there are data gatherers, data processors, data users, and data transmitters. For example, a ship's fire control system may consist of a detection and tracking system, a data processing computer, navigation equipment, a launcher platform, and a missile. From the tracking system, information in the form of video signals is received concerning motions of possible targets, and the navigation equipment gathers information concerning the motion and position of the ship. These two data gatherers send information in the form of A.C. and D.C. synchro voltages to the data processing equipment which is a digital computer. Converters are used to translate the gathered data into data usable by the computing equipment. The data is processed, and information concerning motion and position of the target with respect to the ship is computed. This data is then sent to the launcher platform to select the optimum trajectory path for the missile and inform it as to when to launch. Once the missile is fired, information on target motion is fed to the missile via the tracking radar. Therefore, for compilation of data from a number of sources it is usually advantageous to convert to one language of processing. After processing it is usually necessary to convert again to the language required by the equipment using the processed data. However, converters downgrade a message, which results in a loss of reliability, and also require more equipment, which also results in a reduction of reliability.



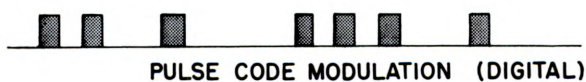
PULSE MODULATION TECHNIQUES

Pulse modulation has become increasingly important in the last few years and is based on one of the following techniques:

- a. PCM, pulse-code modulation
- b. PWM, pulse-width modulation
- c. PTM, pulse-time modulation

In the PTM system, the time interval between two pulses is measured. One pulse is a series of periodically occurring pulses called the synch pulses. The second pulse will occur anytime between two successively occurring synch pulses. Since it is the time at which the pulse reaches a certain amplitude during its rise that is actually measured, the fidelity of the system therefore depends on the ratio of the maximum deviation of the pulse time to the rise time of the pulse. By decreasing the rise time of the pulse, additional bandwidth could be used to improve the signal-to-noise ratio. In the PWM system, the time difference between the rise and decay of a pulse is measured. In this system, the signal-to-noise ratio depends on the accuracy with which the position in time of the pulse edges can be determined. For a given noise level, the error will be proportional to the rise time. In the PCM system, which is digital rather than analog as the first two, we would sample at a rate equivalent to

the synch period in the PTM system and send the samples through a converter for quantizing. The converter then delivers a pulse when the bit is a 1, and no pulse when the bit is 0. In this system, the frequency of errors depends on the signal-to-noise ratio in the channel. However, S/N is usually quite high, and, for all intents and purposes, the frequency of errors is zero. The disadvantage of the PCM system is that degradation of the message may result, due to the quantizing noise which occurs in the analog-to-digital converter.



METHODS OF COMMUNICATION

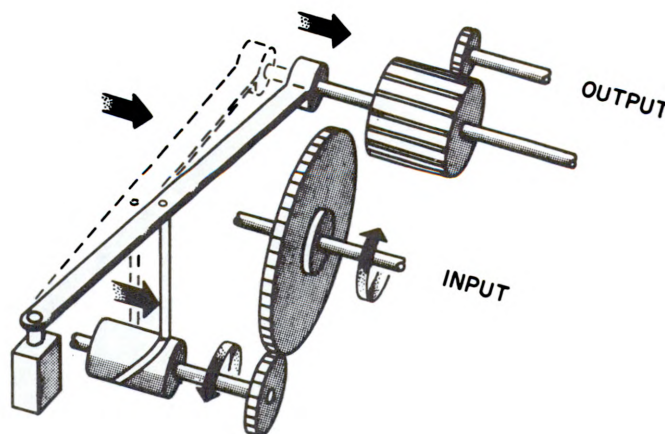
machine-to-machine data links

MECHANICAL COMMUNICATION

Unit-to-unit linkage consists of transmission between machines by mechanical or acoustical means, or light, wire, and radio. Mechanical linkages may be used between two units very close to each other. These linkages are usually in the form of gear trains, manual controls, and belt-driven assemblies. Each linkage is simple, cheap, and secure, providing positive control. The units, though reliable, are limited in speed, capacity, and accuracy. Speed is limited since mechanical linkages possess inertia and in most cases must be geared to the speed of the human factor in the use of these assemblies. Capacity is limited by virtue of the lack of speed. Accuracy is limited by virtue of the drag due to the mass of units and the limitations imposed by space, i.e., only a finite number of teeth can be machined in a gear; any increase in the number of teeth decreases the strength of the piece if the size is not increased.

ACOUSTICAL COMMUNICATION

Communication using acoustics is limited to short ranges. In the audible range, utilizing public address systems affords a very low security, a fair amount of accuracy depending on the system, not as much speed as other systems, and a limited capacity. Audible communication is often used between submarines and ship-to-submarine. Security is excellent when used between ship-to-ship, and visual protection is possible. Underwater audio communication is dangerous, since sound travels considerable distances under these conditions and can be detected. Passive underwater detection



systems have short range, but are advantageous since they do not announce their presence as do active underwater detection systems.

LIGHT COMMUNICATION Communication utilizing light, either visible, infrared, or ultraviolet, provides the maximum in speed on a line-of-sight basis. Infrared and ultraviolet are more advantageous than visible light since they provide better security. Recent developments of lasers, which produce coherent light resulting in high resolving power and narrow beam, show promises of future applications in space communications. However, equipment for utilizing light is expensive, and the same results may be obtained by other methods. Common uses of visible light include blinker lights, flares, and semaphores.

WIRE COMMUNICATION. Communication using wires is perhaps the most widely used. Wire affords good security, speed, capacity, accuracy, and reliability, and is fairly cheap. Examples of the use of wires are the telephone, telegraph, and power transmission. Wire-guided missiles were used by the Germans in World War II for air-to-air interception. In our modern arsenal, the Cobra and SS-10 and SS-11 are wire-guided anti-tank missiles, and the Navy has been using wire-commanded torpedoes for firing from submarines. This type of control and guidance system, wherein the missile is controlled in flight by electronic signals transmitted through wire attached to the missile, is almost immune to enemy countermeasures.



RADIO COMMUNICATION. Radio frequencies used for communications, radar, and other applications vary from 15 dc to more than 30,000 mc. Within this spectrum, groups, or bands, of frequencies have been classified and given arbitrary nomenclature. Each band has characteristics that make it suitable for particular modes of transmission. In the very low frequency and low frequency ranges, the principle mode of propagation is along the surface of the Earth. At these frequencies, reliable communications over several thousand miles may be carried on. However, to effect this transmission, large and powerful antennas and transmitters are required. In addition, VLF can penetrate water to a limited depth, which can be used to advantage in sub-

marine communications and detection systems. In the medium and high frequency ranges, transmission is mainly by waves that travel toward the ionized layers of the atmosphere and are then refracted to the Earth. The use of these frequencies enables communications to be carried on across great distances with relatively small transmitting antennas and low transmitter power requirements. The disadvantage of transmission in these frequencies are mostly environmental, and affect the reliability of the transmission. The time of day, the season of the year, and meteorological and sunspot conditions can cause variations in signal strength and fidelity. Transmission in frequency ranges above 30 mc is essentially confined to straight-line paths (line-of-sight) between transmitter and receiver. The limit on the range of communications when using these frequencies is determined by the curvature of the Earth. This high-frequency band is utilized for narrow-band types of transmission, which are necessary in precision-tracking radar applications.

A disadvantage of radio transmission, since it is in most cases an active system, is that it is subject to countermeasures. Since radio transmission is wireless, it can be intercepted by anyone with a knowledge of the broadcast frequency and the proper equipment. This factor is a disadvantage as far as security is concerned. Once a message is intercepted with the proper equipment, the message can be falsified and retransmitted. Message transmission and reception also can be jammed and distorted. Jamming and distortion of radio transmission can be affected by noise and interference from nearby transmitters operating in the same frequency band. In radar techniques, where detection depends on radio transmission and the return of echoes from a target, the return echoes can be jammed and distorted by various means, such as cluttering the area with bundles of materials which are opaque to radio transmission waves. The reflected false returns act as decoys and can obscure the real target. A second method is to determine the frequency at which a transmitter is broadcasting, and generate return signals that can be mistaken for the actual target. There are, of course, methods of countering these jamming techniques, i.e., counter-countermeasures.

FREQUENCY RANGE	WAVE LENGTH RANGE	CLASSIFICATION	
10kc - 30kc	30,000 to 10,000 meters	VLF	Very low frequency
30kc - 300kc	10,000 to 1,000 meters	LF	low frequency
300kc - 3000kc	1,000 to 100 meters	MF	mid frequency
3mc to 30mc	100 to 10 meters	HF	high frequency
30mc to 300mc	10 to 1 meters	VHF	very high frequency
300mc to 3000mc	100 to 10 cm	UHF	ultra high frequency
3000mc to 30,000mc	.10 to 1 cm	SHF	super high frequency

RADIO FREQUENCIES

MODES OF INTELLIGENCE TRANSMISSION

All modes of intelligence transmission which are in common use and employ carriers or sidebands can be basically expressed as variations of either carrier amplitude or carrier frequency. In common use are three basic systems of amplitude modulation and one of frequency modulation. In addition, there are several combinations of both in use. An example of the latter would be the broadcast of a stereo program over both an AM and an FM radio broadcast band. In this example of combined usage, each modulator is used separately to transmit one channel of information. Another aspect of a combined use would be in an interrelated

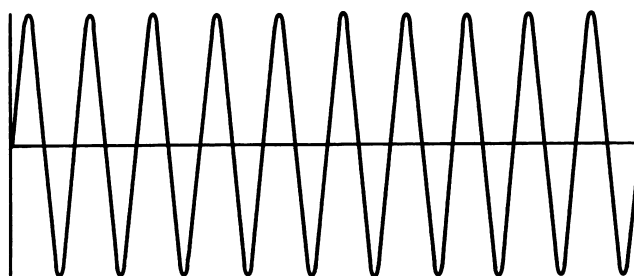
transmission where one modulation becomes a carrier for a new modulation. This type of system is usually referred to as a subcarrier system.

amplitude modulation

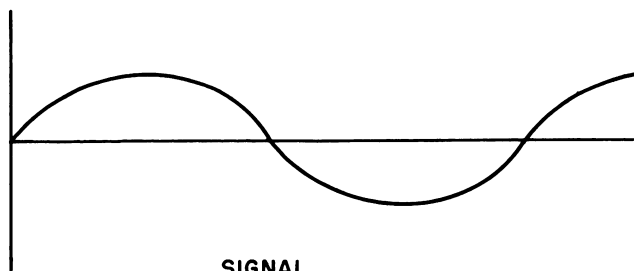
In amplitude modulation, the three basic systems are:

- Carrier and two sidebands
- Suppressed carrier, two sidebands
- Single sideband, suppressed carrier

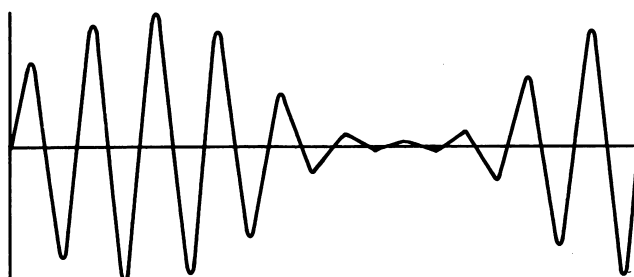
The basic system of a carrier and two sidebands is the classic example of a modulation system wherein a carrier sine wave is amplitude modulated by a signal sine wave.



CARRIER



SIGNAL



AMPLITUDE MODULATED WAVE

The combination of the carrier sine wave and the signal sine wave produces three unique frequencies: the frequency of the carrier, and one upper and one lower frequency differing from the carrier frequency by an amount equal to the signal frequency. For example, if the carrier is 100,000 cps and the signal is 10,000 cps, the modulated carrier will contain three components:

- 100,000 cps (carrier)
- 110,000 cps (carrier + signal frequency)
- 90,000 cps (carrier - signal frequency)

The sum and difference frequency components are called the sidebands. These sidebands frequencies carry all the information the original modulating signal contained. In amplitude modulation the sinusoidal carrier current may be represented by the expression:

$$i_c = I_c \sin w_1 t.$$

If the carrier is modulated by an original of value m ,

$$i_m = m I_c \sin w_2 t,$$

the resulting modulated current will be:

$$i = I_c (1 + m \sin w_2 t) \sin w_1 t.$$

The expansion of this expression and substitution of equivalent trigonometric terms for $\sin w_1 t \sin w_2 t$ gives an expression for the modulated current:

$$i = I_c \sin w_1 t + \frac{m}{2} I_c \cos (w_1 - w_2) t - \frac{m}{2} I_c \cos (w_1 + w_2) t.$$

This expression shows that the amplitude-modulated carrier is made up of three components. The carrier, in addition to aiding the transmission of the signal sine wave, is also necessary in the detection process, since the carrier and at least one sideband must combine in a nonlinear device to cause the demodulation of the carrier wave. In most cases, after use in the detection process, the carrier is discarded as no longer needed, and the information in the sidebands is retained. Since the carrier is necessary only at the receiver end or termination of the communication circuit, it is frequently discarded at the transmitter. This leads to a consideration of the suppressed-carrier, two-sideband, amplitude-modulated mode of transmission, the carrier is not employed, but is suppressed. However, the carrier must be supplied at the detector so that detection of the signal sine wave can be accomplished. The detector must be supplied with a local carrier identical to the original suppressed carrier in waveform, frequency, and phase. Once this is done, the intelligence contained in the sidebands can be regained. If this is not done, then the demodulated information will be distorted.

The third basic amplitude-modulated system employs the suppressed carrier with a single sideband. This system has advantages over the aforementioned methods in that a reduction in power and bandwidth can be realized. By the transmission of only one sideband, the bandwidth required to transmit all the information possible is reduced by a factor greater than two.

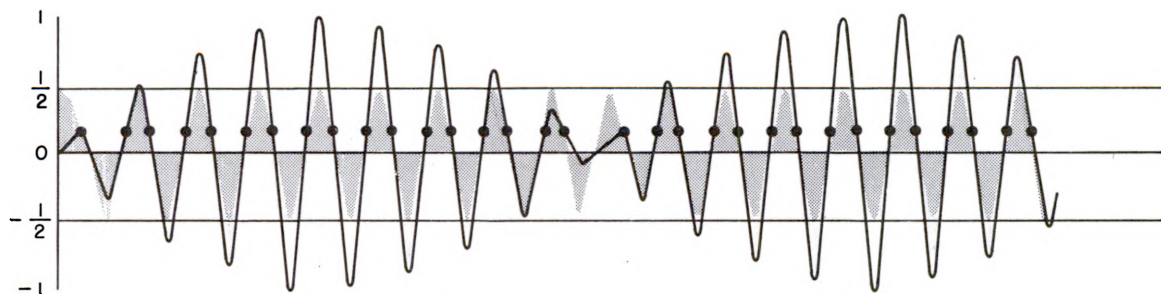
With all signals there is associated some noise from a particular noise source. This noise power may be expressed as the product of the noise power per cycle per second, and bandwidth in cycles per second, or $N = wB$, where N is the noise power, w is the noise power per cycle per second of bandwidth, and B is the bandwidth in cycles per second.

Therefore, the noise power is directly proportional to the bandwidth. In both cases of suppressed carrier systems, the entire power capacity of the system can be used safely to generate or transmit sidebands. In the single-sideband system, the signal power can be expressed as S , and in the double-sideband system the signal power in each sideband is expressed as $S/2$. In the first system, that of carrier and double-sideband, assuming the signal power for this system is equal to the signal power in the other systems, the signal power for each sideband is $S/6$ and for the carrier is $S/1.5$. The following table is a summary of amplitude-modulated systems with regard to the bandwidth and ultimate signal-to-noise ratios.

Amplitude-Modulated System Characteristics

System	Double Sideband and Carrier	Double Sideband Suppressed Carrier	Single Sideband Suppressed Carrier
Sideband Power	$S/3$	$S/2$	S
Carrier Power	$S/1.5$	0	0
Transmission Bandwidth	$\approx 2B$	$\approx 2B$	B
S/N Ratio	$\approx 1/6$	$1/6$	1

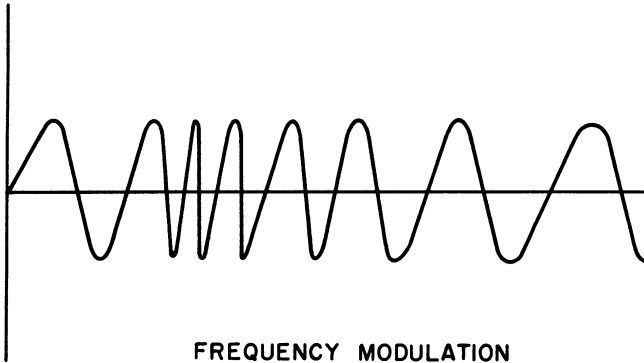
Of the three systems discussed above, the suppressed-carrier single-sideband system is most advantageous when considering effective employment of bandwidth and system power capability.



GRAPHIC EXTRACTION OF UPPER SIDEBAND

frequency modulation

Frequency modulation refers to the modulation of the frequency of a carrier wave. This type of modulation is often referred to as phase modulation; however, there are basic differences between frequency and phase modulation. These differences can be demonstrated by the following discussion of an A.C. Generator.



If the speed of the alternator is varied with time, and if simultaneously the field of the alternator is also varied in a manner to maintain a constant maximum voltage, then the resulting frequency of the output must vary even though the amplitude of the waves remains constant. This is analogous to frequency modulation. Now, if the speed and the field of an alternator are kept constant, and if the stationary part of the alternator is movable so that it can be oscillated back and forth through a small angle, the resulting wave form of the alternator represents phase modulation.

Applications of frequency modulation occur in detection systems using the Doppler principle for detection of moving targets, and in telemetering systems analyzing the performance of missiles under test.

Unlike amplitude modulation wherein the bandwidth is dependent on the frequency of the modulating signal, the bandwidth in frequency modulation is proportional to the bandwidth and amplitude of the modulating signal. These two parameters are further related by a third parameter referred to as the modulation index:

$$\text{Modulation index} = \frac{\text{Carrier frequency deviation}}{\text{Modulating frequency}}$$

Modulation of the frequency of the carrier wave occurs at a rate which is proportional to the frequency of the modulating signal, and the maximum deviations of the carrier frequency are proportional to the amplitude of the modulating signal. Bandwidth (B) of the modulated wave is demonstrated by:

$$B = c_1 \Delta f_1 + c_2 f_2,$$

where Δf_1 = carrier frequency deviation

f_2 = modulating frequency

and c_1 and c_2 = constants.

Letting $m = \Delta f_1 / f_2$ as stated above, then:

$$B = c_1 m f_2 + c_2 f_2 = (c_1 m + c_2) f_2$$

where m is the modulation index.

Constants c_1 and c_2 depend upon the necessity for reducing the large actual bandwidth to some finite bandwidth. In FM broadcast practice, these constants are selected so that the variation of the carrier from its mean value is ± 75 kc. If this procedure were not followed, the maximum deviation in the negative direction could conceivably swing the carrier from its normal value to zero. In a similar manner, a positive deviation could push the carrier frequency to several times its normal value. In most cases, however, practical circuit limitations dictate a deviation of only a small percent of the carrier if severe distortion due to nonlinearity in the r-f circuits is to be avoided. One of the most valuable properties of frequency modulated systems is the improvement in signal-to-noise ratio as compared with amplitude-modulated systems. The noise voltages encountered in both systems can be considered to be modulated both in frequency and in amplitude. In amplitude modulation, the disturbances are superimposed upon the signal voltage, and thus these disturbances appear in the output of the detector and finally in the audio output. In frequency modulation, the amplitude disturbances are eliminated by the use of a discriminator that tends to be unresponsive to amplitude variations, and, when used in conjunction with limiters, amplitude disturbances can be almost completely eliminated. The frequency-modulated component of the noise can be negated by employing a large frequency swing of the transmitted signal. In this manner, the signal can be made as large as necessary in proportion to the noise signal, resulting in a desired signal-to-noise ratio in the output.

In comparing the signal-to-noise ratio in frequency modulation to the signal-to-noise ratio in singleband amplitude modulation, a relationship between the two has been found to exist. This relationship, expressed as

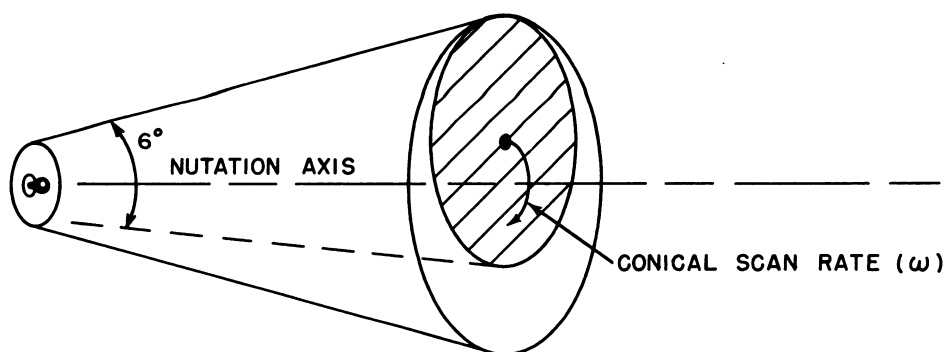
$$B_{FM}/B_{AM} = 8/3,$$

accounts for the higher quality of reception inherent in an FM system in comparison with the finest AM system. Uses of frequency modulation in missile guidance and control take on a number of forms, two of which are intentional frequency modulation and natural frequency modulation. The former is utilized in the standard FM altimeter in controlling the altitude of a missile. The latter form, called the Doppler principle is used in the detection of moving targets.

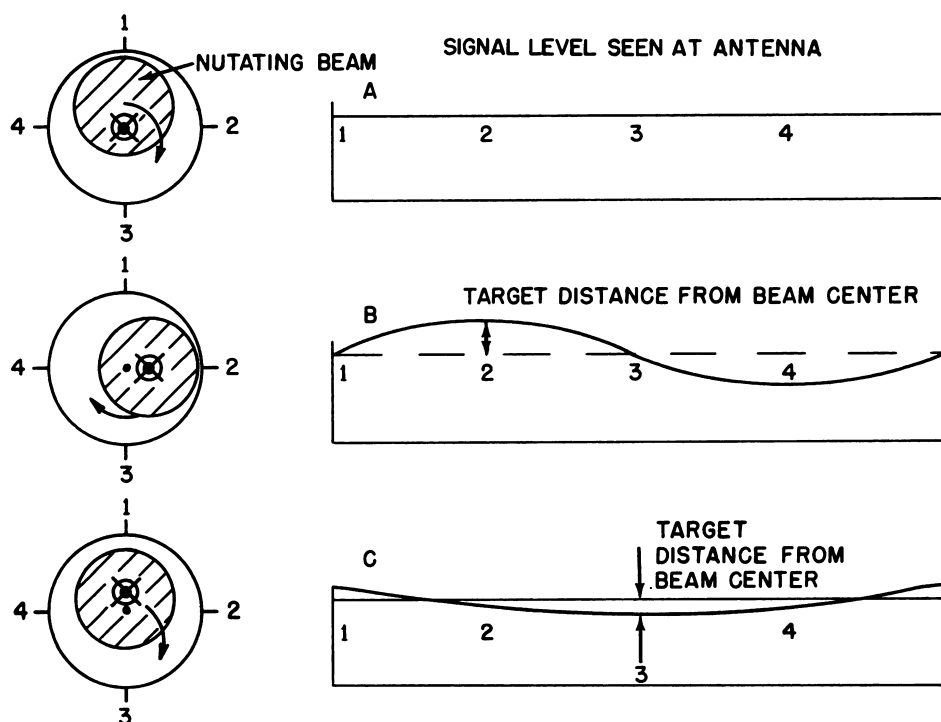
subcarriers

Subcarrier systems utilize one modulating signal as the carrier wave for another system. For example, consider a fire control radar system. In this system, an r-f carrier is amplitude modulated by a series of short pulses. The modulated carrier is transmitted and then reflected from a target. These reflected pulses are then modulated by the range to the target.

A more concrete example follows. In a scanning type of fire control radar, the center line of the antenna beam is required to describe a cone in space by nutation of the antenna feed system at some angular velocity, ω . When the target lies in the center of the beam, the signal received by the radar will be an r-f carrier, the



FORMATION OF CONE BY NUTATION OF ANTENNA BEAM



TARGET POSITION IN BEAM VS SIGNAL AMPLITUDE

amplitude of which will be pulse modulated. If the target moves off the center beam for distance x , the signals received by the radar will be similar, but further amplitude modulated by scan frequency ω . The amount of amplitude modulation is proportional to distance x from the center of the beam. In addition, the phase of ω is proportional to the angular direction of point x referred to some reference line. Further, if the target at x is moving in a sinusoidal manner, then ω would become the carrier for a modulating signal proportional to the amplitude of the sinusoid about point x . The phase modulation of ω indicates the direction of target motion. The spectrum received by the radar is then composed of many subcarriers each of which contains valuable information concerning the position and motion of the target. Therefore, each of these subcarriers must be processed so as to obtain all this information.

GUIDANCE INTELLIGENCE The nature of guidance intelligence, initially, is the setting down of a clear statement of the objective of a missile system and the defining of the limits of its operation. Secondly, once the objective is formulated and the limits known, the accuracy of the guidance system depends on the type of information received from the target by the detection devices used in the weapon system. Third, a suitable guidance system, once target characteristics and missile performance parameters are known, can be readily determined. Fourth, once the general form of the guidance system is determined, it is possible to investigate some basic facts which, in the processing of the information, are important to the outcome. It is at this stage of development that information bandwidths are determined. At this point, the transmission of the guidance intelligence between computer and missile becomes all important.

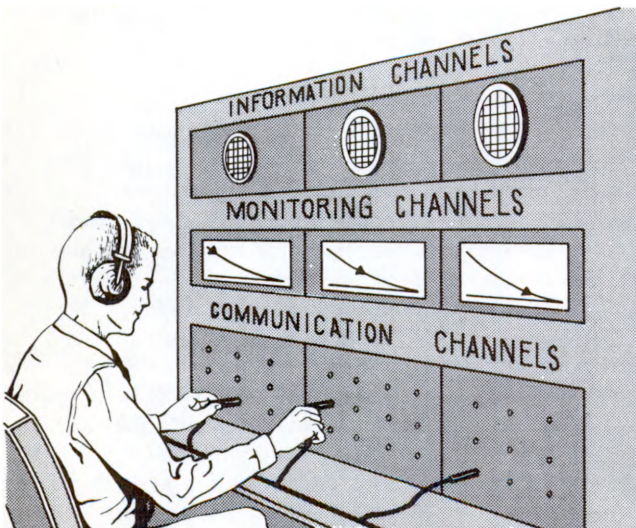
MAN-MACHINE COMMUNICATIONS

Communications between man and machine are concerned primarily with the need for inputs and outputs both of which are compatible with man and machine. Since the characteristics of both man and machine differ, there is need for specific data links which optimize the combination of both man and machine in a system. The primary need is for short-range data links, that is, man must be positioned as close to the machine as possible. This facilitates correlation of the input data generated by the man and inserted into the machine. In this instance, the machine informs the man exactly what it has received, and, by doing so, checks the accuracy of the man's input and acknowledges the receipt of same.

In the output flow of information, the machine must provide the man with devices sufficient for understanding the solutions generated by the machine. These outputs must not be such as to overload man's ability to assimilate all the output data, for, if they do, then the man-machine combination is not performing at maximum efficiency.

system outputs

The machine outputs are of two forms, material and information. A material output, that is, an output which affects some action, is not of primary concern in a discussion of the man-machine combination. The output of information however, is directly associated with the man-machine combination. This output is usually in the form of a display which allows man to make decisions based on the information on the display. The machine must be discriminate in the rate of presentation and the amount of information present at any one time. If this discrimination is not adhered to, the effectiveness of the man's decision is greatly reduced. Methods and types of machine outputs are either tabular or pictorial, and either temporary or permanent.



information for human operators and monitors

Aside from the machine outputs containing information for decision making, there are other outputs which provide data for the operator and for monitoring the efficiency and performance of the machine. The data for the operator is in a form of positive notification that the inputs have been received and understood, that the command has been carried out, and that the action has been taken. A simple example is the ringing heard by a telephone user after dialing a number. This is not the same ring heard by the recipient of the call, but rather is a signal generated somewhere else in the telephone system. The ringing heard by the caller merely informs him that he has made a connection and must wait until the recipient answers. If the signal were not present, the caller would become impatient and hang up before the recipient had a chance to answer.

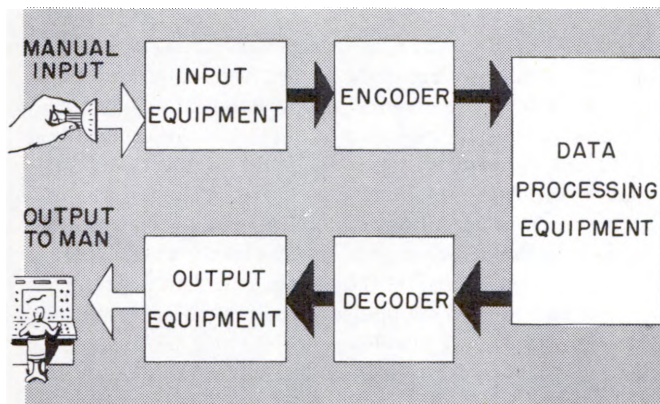
In today's complex systems, there is need for monitoring the performance of the machine during operation. Therefore, specific displays of the system give performance indications which are noted by the responsible maintenance personnel. These results are useful both as inputs for future tests and as raw data in further studies for the improvement of the system.

system inputs

The inputs of a system can be classified as follows: information at hand and information at a distance. The prime concern here is information at hand. This information is usually supplied by a human operator, and must undergo some form of translation from the language of the operator to the language of the machine. In many automatic systems, the inputs are in a form which remains static throughout the operation of the machine; for example, the numerous programs that are inserted into a computer. The program is generated by the man before insertion into the machine. Once the machine commences operation, the only inputs it then receives are from memory sources. Another example of the machine receiving only initial inputs is in an automatic tracking system. In this type of system, the radar unit has a long-range search unit and an automatic tracking unit. While the search unit detects the presence of targets in a particular segment of the search area, the in-close tracking unit is activated only on command of the operator to track a desired or chosen target. In this case, the human operator selects a desired target, and furnishes this initial bit of information input to the automatic tracking system, which then takes over the actual tracking of the designated target. However, should a malfunction occur, the system should be so designed that the man can assume manual operation of the unit. In some automatic machines the only function of the operator is to set up the machine for a particular operation. In many instances, a system is designed to operate on the average rate and number of inputs. That is, the system will respond in the desired fashion to events of high probability and is designed to favor the most

likely. This is statistical in nature, and normally no problems arise. However, if an event occurs, that is low in probability and the machine can not handle it, then the human operator must take over. Thus, in general, there is always a need for mechanical override. In other machines, the human operator must make the decisions for inserting orders into a machine. For instance, a system designed to engage high performance aircraft may not be able to handle a PT boat or a ship target. In this event, man must take over and provide control for this low probability situation.

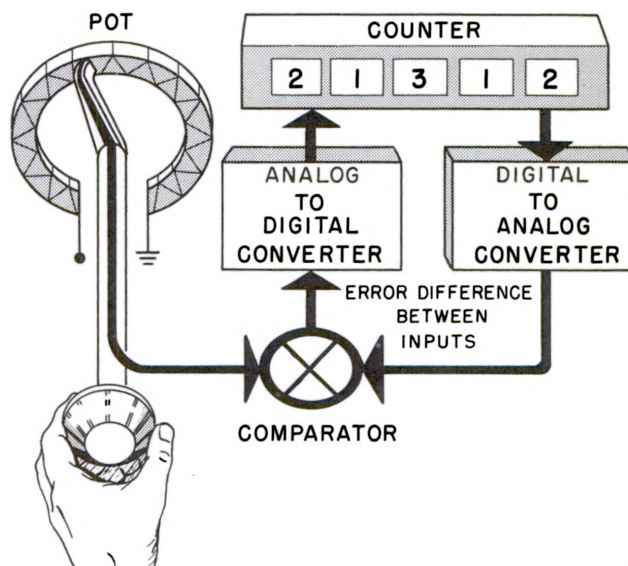
All inputs to the machine must be standardized, that is, inserted into the machine in a form compatible with the machine. Equipments used for the insertion of standardized information may be either analog or digital. Analog equipment includes the pantograph, the joystick, and the wheel or dial. These equipments can be used singularly or in combination depending on the ability of the operator to handle them. The common digital equipment are dials and keysets. Keysets can be of a specialized or general purpose type.



man analog by nature

The human being, at least in his input and output, is inherently analog, and so the machine with which he deals must also be analog in most cases. Of course, the entire machine does not have to be analog, but only the input and output equipment. The intermediate equipment (data processing equipment) can be digital. If this is the case, then there is a need for converters, both analog-to-digital and digital-to-analog. The analog-to-digital converter, often referred to as an encoder, effects the translation, in most cases, from discrete voltages to digital numbers. In other cases, the encoder operates directly from some analog other than voltage. In the former instance, man's input may be to rotate a dial to a particular position. The dial is mechanically coupled to a wiper arm of a potentiometer which picks off a voltage equivalent to the position of the dial. This voltage is fed to a comparator unit designed to measure the input analog voltage and a voltage proportional to the binary count in a counter. If the comparison is not favorable, the comparator emits count pulses which advance the binary count in the counter. The process continues

until the analog voltage equals the proportional voltage, at which time the binary output is the equivalent to the dial setting. The digital-to-analog converter, often referred to as a decoder, usually translates digital quantities into voltages, since voltages may be easily converted to other analogs as desired. In converting from digital-to-analog, one is concerned with the serial or parallel form of the digital number. In decoding serially, switching and timing requirements greatly complicate the equipment. To decode parallel form digits, the equipment is relatively simple, and in common usage one finds the so-called transistor-operated voltage divider or TOVD.



NEED FOR BANDWIDTH MATCHING

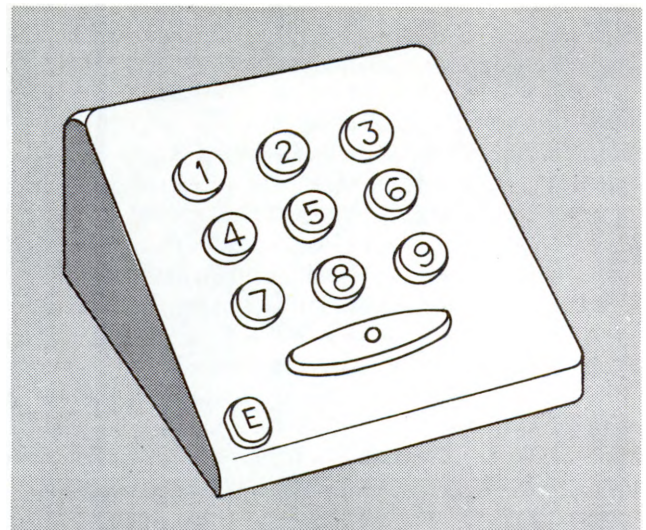
Maximizing the optimum utilization of man and machine comes about with proper matching. Without this match, man would not be able to keep in step with the machine, and, as a consequence, the efficiency and worth of the machine would decrease. Man can give the machine directions in the form of orders at a specific rate. If the machine spends most of its time just awaiting these orders, then the cost and design characteristics which went into the development of the machine are wasted for all intents and purposes. The machine must be designed for utilization of man's bandwidth. That is, man's limited bandwidth must be taken into consideration when determining how fast a machine must be in acting on an input, and how fast and how much information the machine must deliver to its output devices. Conceivably, a machine with good design qualities could deliver too much information. Whenever there is too much information for a man to digest, there are four steps which may be taken: a) eliminate some of the information which is unimportant and non-pertinent, b) condense some of the information by summarization, c) classify the information so that it may be called up by category when needed, but not otherwise and d) divide the responsibility for decision making between two or more people in such a way that none of them needs all the information.

CONTROLS AND DISPLAYS

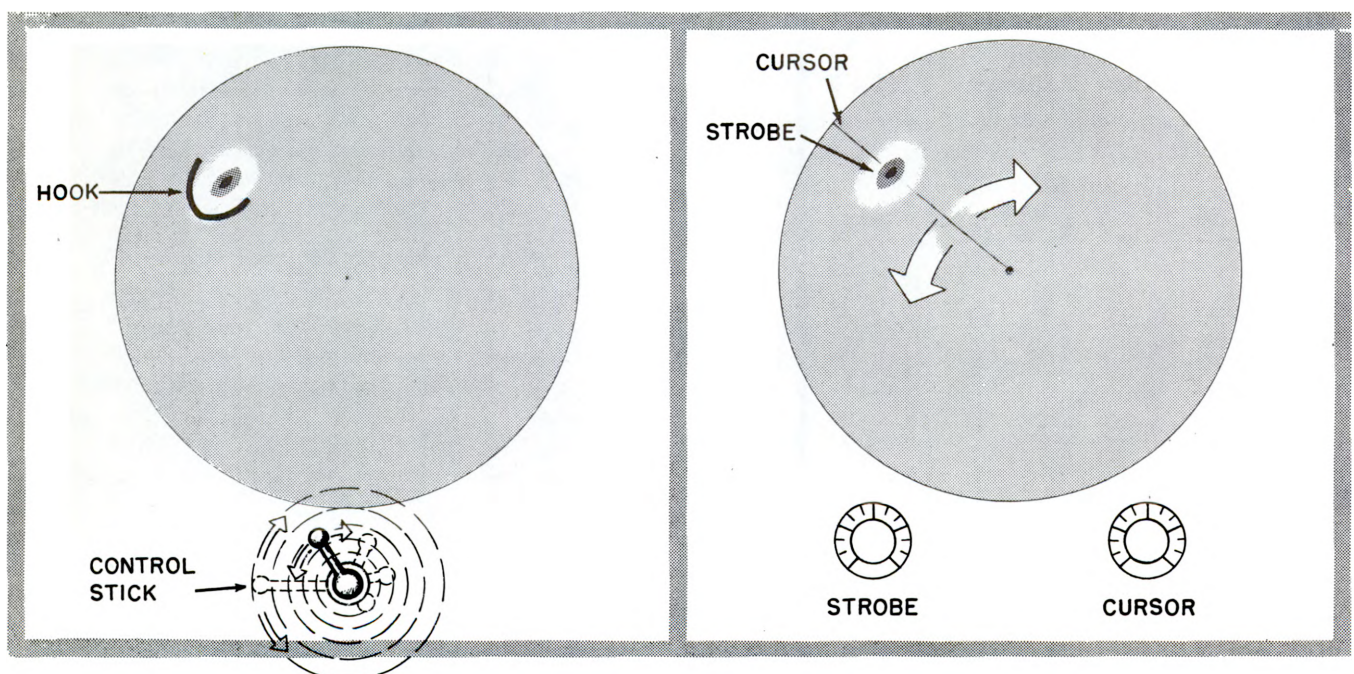
controls

Controls for man's inputs, as noted previously, are either analog or digital in nature. The pantograph, which is an analog control, can be moved left-right and forward-backward, while constrained in a plane. The joystick can move left-right and forward-backward, while one end is contained in a swivel joint. The wheel or dial can rotate clockwise or counterclockwise. In many aircraft, the joystick and the wheel are combined to give three degrees of freedom: left-right, forward-backward, and rotational. One system utilizes a joystick to center a hook, displayed on a PPI scope, about a target blip. Once the hook is centered about the target, the target parameters can be entered into the system by depressing the entry pushbutton.

SPECIALIZED KEYSET Many radar systems use dials or wheels to center strobes and/or cursors over a target on which information is desired. However, the disadvantage of this method as compared to the joystick method is obvious since the operator must manipulate two wheels, one for the cursor, and one for the strobe. In the previously stated example, one control was necessary to specify the position of the target. The specialized keyset, which is a digital control, is designed with as many columns of keys as there are categories of information and with sufficient rows in each column to insert the appropriate digit for that category, for example, the cash register. In the smallest cash registers, there are two columns for cents and one or two columns for dollars. Each column contains digits 0 through 9. This particular register has a capacity for entering from 0 to \$99.99. Usually each column or group of columns is distinguished from the others by color, shape, and/or size.



GENERALIZED KEYSET. The generalized keyset has ten keys representing digits 1 through 0, and other keys for execute and reset. Depression of any key inserts a number in the order of its depression into the keyboard and produces the sequence of digits in a window which the operator observes. As soon as the information is complete, the operator compares the number in the window with the desired number he wished to insert, and, if they agree, he then depresses the execute key. This operation sends the information into the system. It should be noted that, with each control discussed, there is an associated display which indicates to the operator the results of his actions.



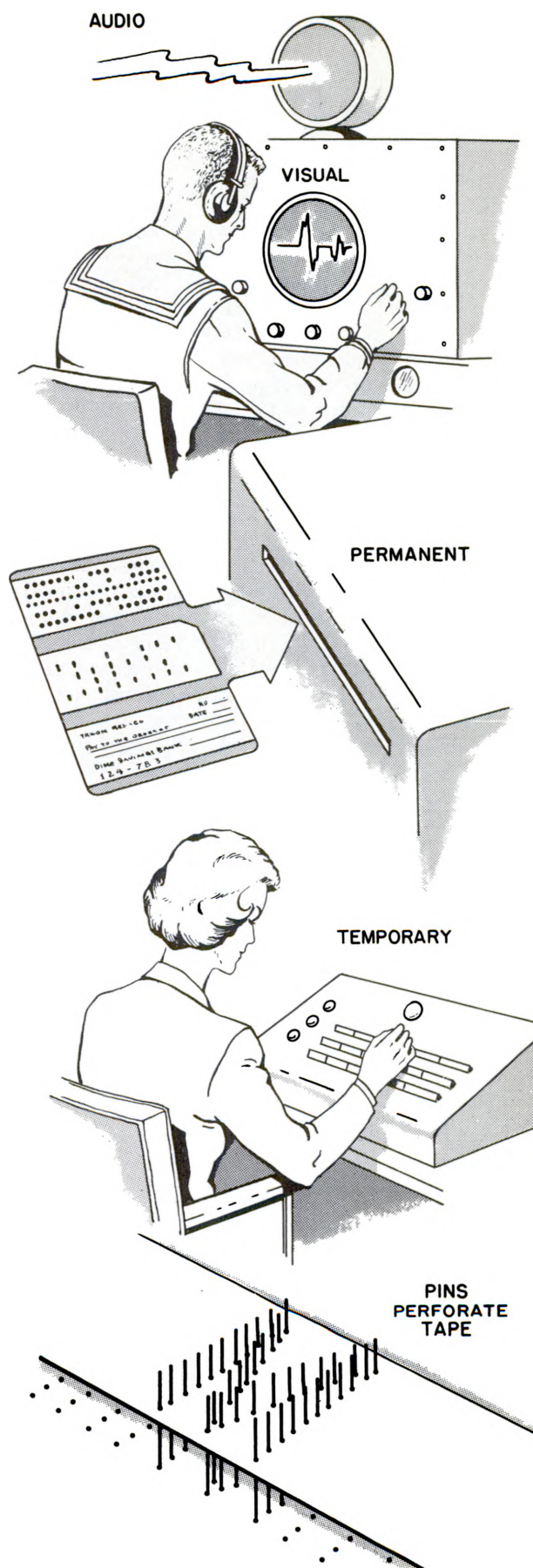
displays

Machine displays consist primarily of two types: visual and auditory. Visual displays may be classified as follows: tabular or pictorial, and temporary or permanent. Tabular displays present conventional symbols, such as numbers, or ordinary writing in the form of letters or words, while pictorial displays present a wide variety of other techniques such as graphs, maps, etc. Typical tabular material can be presented by pictorial techniques, but pictorial material can not be presented by tabular techniques. Permanent displays are recorded on paper or photographic film, while temporary displays are erasable.

Documents such as maps, graphs, photographs, etc., are classified as permanent pictorial displays. Techniques for the rapid production of permanent pictorial displays include the recently developed Land-Poloroid photograph, which produces a permanent photograph in less than a minute, and xerography, which produces permanent records by a new chemical process known as dry photography. Some of these techniques use chemicals, and others use photoelectric or photoconductive phenomena known as electrophotographic techniques. Radio facsimile is a technique for the transmission of pictorial information to distant points and the reproduction of the original at the receiving terminal for a permanent record. The sensitivity and the fidelity of the received display are low in comparison with other photographic display devices, but have the advantage of being transmitted long distances.

Permanent tabular displays include techniques known as telantograph, a mechanical pencil which writes on a moving roll of tape. A typical device is the teletype, which employs a particular code for ordinary letters and numbers which are transmitted at high speeds and decoded in tabular form at the receiving terminal. The cathode-ray tube is utilized as a temporary pictorial display, as in the oscilloscope; grids or other references can be permanently or temporarily overlaid on the front of the picture tube. Some recently developed tabular displays utilize beam-switching tubes or cathode-ray tubes to display visual solutions to solved problems. Prearranged degrees of beam deflection can also actuate relay systems to perform mechanical operations such as the formation of electronic characters as illustrated. The difference is that instead of light that is turned on or off or moved to various positions, tiny pins are mated to a carbon ribbon to make a dot. The grid of these dots covers each letter space. An electrical signal causes the desired pattern of dots to press against the carbon, resulting in a character formation.

Each of the aforementioned methods of display are visual by nature. In addition, there are acoustical displays, usually in the form of speaker outputs. A typical system using acoustical methods is sonar detection. In this system, an experienced operator can distinguish between sounds caused by types and sizes of underwater craft, marine life, and other underwater phenomena.



Human engineering is the study of the role of the human operator in an automatic or semi-automatic system. The primary consideration is the maximum effectiveness of the system.

The first task is to determine those areas in which a man's capabilities exceed those of the machine, and those areas in which he is surpassed by the machine. Once these areas are formulated, the man is assigned specific tasks and the machine other tasks. The second step is to determine the best working combination of these tasks so that the entire operation is complimented by the use of man and machine. From this standpoint, the machine may be designed to give aid to the man in the performance of his tasks. This serves to decrease human effort, and, therefore, increases human performance. If the man is operating a control and monitoring the effects of the control at the same time, the machine can aid the man by indicating the results of the man's manipulation of the control. The third step is the study and design of controls and displays for use by the man. In this area the physiology of man is important. An exact knowledge of the characteristics of the five senses is required. The study of the limitations and capabilities of these senses and the effects of fatigue go a long way in insuring optimum compatibility between man and machine. The fourth step is a final evaluation of the man-machine system. In this study, it is all important to determine the operating conditions for both man and machine, the loads placed on the man during various phases of the operation, the effects which occur at breakdown as far as the overall system is concerned, and the success of the man-machine combination.

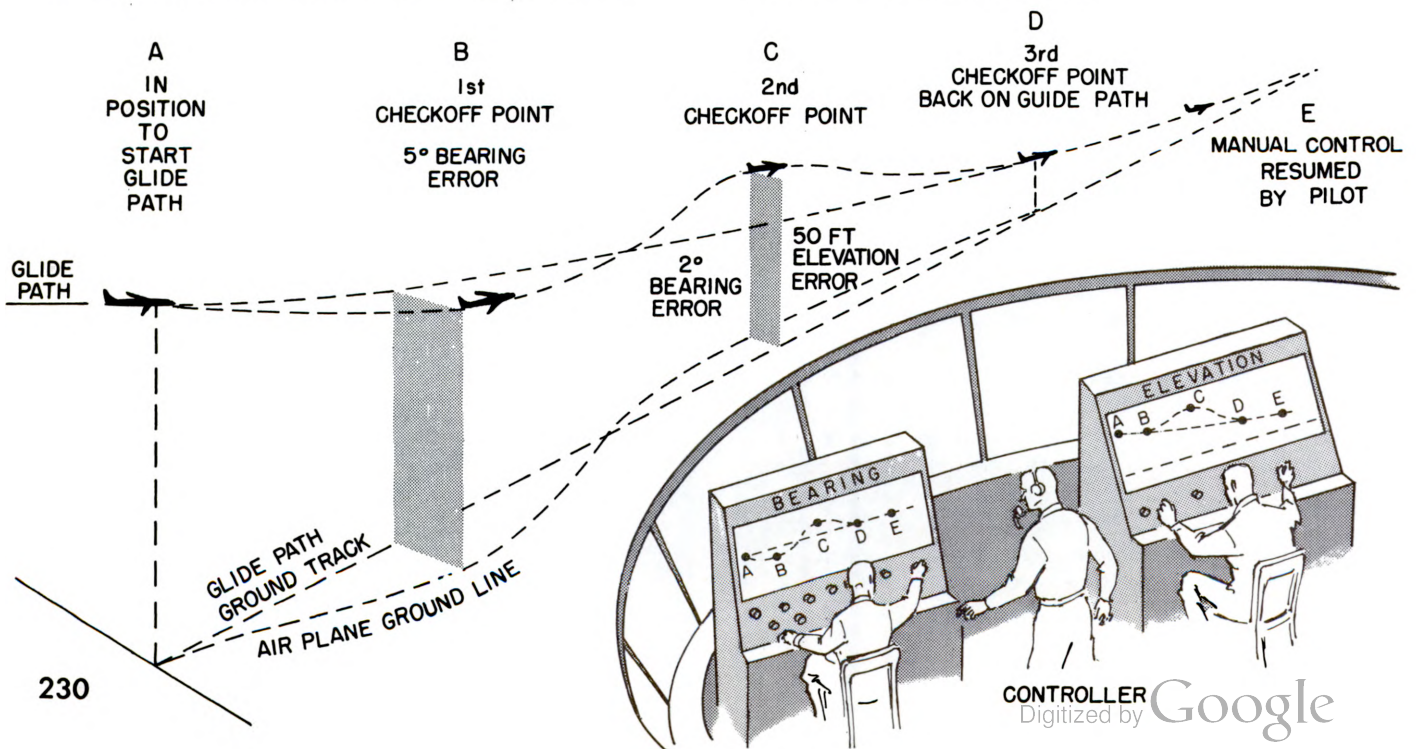
degree of human participation

The degree of human participation in the man-machine system is of utmost importance. Should the man be relegated merely to monitoring, checking, and maintaining the machine; should he be assigned the task of primary control, using the machine as an assistant; or should he be assigned some task between these two extremes?

Before these questions can be solved, the human engineer must evaluate the system in the following terms:

- a. Overall Requirements
 - (1) Minimum performance
 - (2) General environment
 - (3) Life of equipment
 - (4) Size
 - (5) Versatility
- b. External Environment
 - (1) Signal input and output
 - (2) Limit on performance
- c. Internal Environment
 - (1) Component life
 - (2) Space
 - (3) Vibration
- d. Data Requirements
 - (1) Arithmetic
 - (2) Conversions
 - (3) Decisions
 - (4) Interface
- e. Outputs
 - (1) Type
 - (2) Amount

Once this evaluation is complete, the human engineer can assign the responsibility for these functions to man or machine depending on which is the better performer. As an example, a large New York City bank requires its customers to fill out a slip of paper requesting a withdrawal or deposit. The customer usually writes the necessary information on the slip in longhand. After the transaction is made, the paper is given to a keypunch operator, who punches holes in a card reflecting the information on the paper. The punched card is then fed into the machine. There is little doubt that the machine could not perform the task of the keypunch operator since the slip was handwritten.



HUMAN ENGINEERING

On the other hand, regarding arithmetic computation, the machine is an excellent and very rapid computer, whereas the man is a relatively slow and poor computer. There is a machine which has been programmed to calculate the value of π to 100,000 decimal places. This calculation required approximately eight hours to perform. It would take years for a man to accomplish this task.

reliability and efficiency of the system

The human engineer must evaluate man's potential and the machine's ability to do the work and then integrate these two factors to obtain an optimum system. This essentially boils down to a question of how the man can check on the machine and how the machine can check on the man. One approach is to duplicate the man, i.e., have two men, one an operator and one a verifier, both of whom enter the same data at the same time into the machine. The machine is designed to accept only identical data from both operator and verifier. This duplication reduces human error to a great extent and can be expanded simply by using more operators and more verifiers. The probability of both operators making an error at the same time is remote. This method is employed in the Bomarc missile system.

Another approach would be to have the machine aid the man. A closed loop tracking system is a typical application. The man controls the machine until it locks on the target. Then the machine, by proper calculations of the rate of speed and direction, can predict the location of the target at any future time. This aids the man in that once he locks on, he will not lose the target and can perform other tasks such as the determination of the characteristics of the target and how to engage the target most effectively.

PREDETERMINED DISPLAY INFORMATION

A basic requirement of a display unit is that a man must have information regarding the effects of his control movements. A factor to consider in this regard is the requirement for previous knowledge of an effect which a control movement will have. The system should contain a single display unit which will give the operator continuous data on the results of his actions before the results become available from the system's indicating devices. This process is known as "quickenings", and its effect is to tell the operator the effectiveness of his control action during the operation. An example would be the glide path location indicator in a ground control approach (GCA) radar set, which allows the operator continuously to see the result of his control effort.

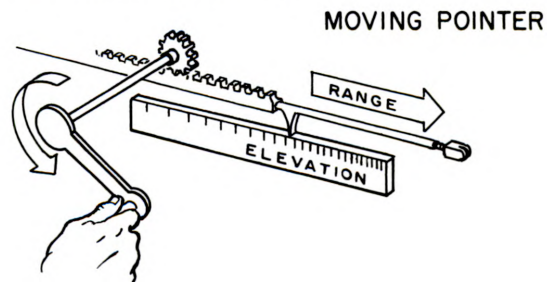
Since man repairs the machine, he must be aware of a fault and have some idea of how to locate the malfunction. To facilitate this, the machine must provide indications of faults, and the design of the machine must take into account the problems of the maintenance man.

man's physical capabilities

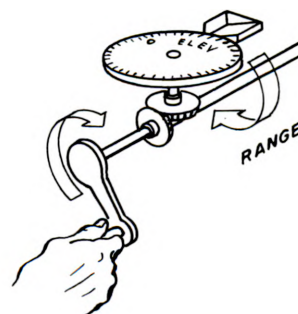
In the design of systems components, the human engineer must have appreciable knowledge of man's physical limitations and capabilities regarding the five senses and human fatigue. With this knowledge, the human engineer can design controls and displays which will enable or allow man to perform his tasks with the least amount of error and the greatest amount of proficiency.

In general, man is required visually to monitor numerous types of displays, listen for various sounds, manipulate innumerable types of controls, and give verbal directions. All these actions may occur simultaneously, in rapid succession, or one at a time at a relatively slow pace. All these factors are to be considered by the human engineer.

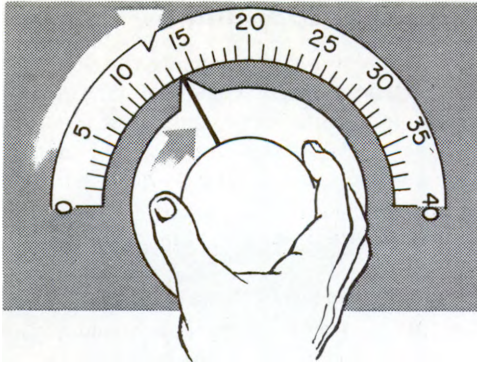
In the design of displays, the question must be asked, "To what use is the display going to be put?" Four possible answers are: quantitative reading, check reading, setting, or tracking. Once this question is answered, the next is, "What type of display would be best suited - moving pointer, moving scale, or counter?" For quantitative reading, obviously it would be the counter; for check reading, the moving pointer; for setting, either the counter or the moving pointer; and for tracking, the moving pointer. It is noted that the moving scale display, although useful for quantitative reading, setting, and tracking, has proven to be a poor indicator from a standpoint of human engineering.



MOVING SCALE



RANGE IN MILES	ELEVATION IN FEET	BEARING IN DEGREES
0 0 4 0	7 5 6 1 2	3 1 4



In the design of displays, the following three principles should be adhered to:

- a. A clockwise rotation should increase the value of whatever is controlled.
- b. The knob and the scale should move in the same direction.
- c. The numerals on the scale should increase in a clockwise direction.

The human engineer must determine whether an operator can detect an auditory signal such as a warning sound, and understand tone-coded or time-coded sounds in the presence of noise, interference or deliberate countermeasures. Female voices are utilized in aircraft warning systems because of clarity of tone and ease of understanding.

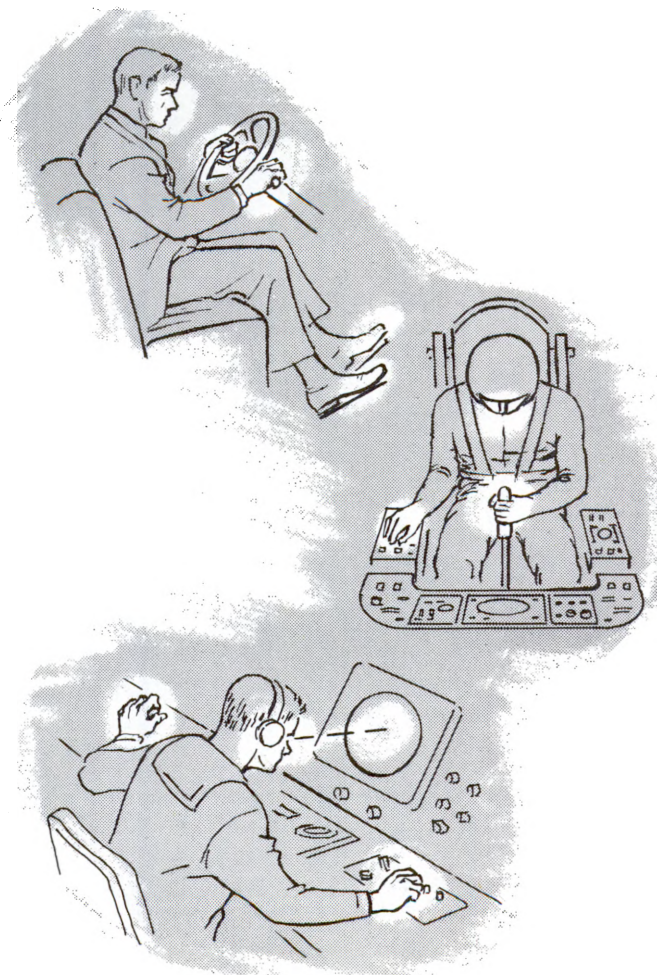
In considering the operator's need for manipulating controls, the human engineer concerns himself with man's sense of touch. This sense is composed of many different senses, i.e., the sensations of hot and cold, tactile discrimination, and others. The tactile sense which is associated with feel by the finger tips, is of particular interest to the human engineer. He is also interested in the kimesthetic sense, which allows us to know the position of our extremities even when we can not see them. These areas are studied by the human engineer, since, in many cases, an operator will be using controls at which he will not actually look. The parameters which must be taken into consideration in the design of controls are as follows: the position of the control with respect to the man, the distance between the controls on a single control panel, the color of the controls (in the case where a man may be able to glance at them), and the shape of the controls. All these parameters have been studied experimentally with the expectation that, once all factors are known, the optimum arrangement and design of the controls can be made. A principal concern of the human engineer in the design of controls covers the area of motion utilized in the manipulation of the controls. There are two types of motion: fixed and ballistic. In fixed movements, one set of muscles works against another, resulting in a tense, tightly controlled movement. In ballistic movements, one set of muscles does all the work, while the opposing muscles are completely relaxed. However, studies of these movements have not resulted in a clear understanding of the advantages which this knowledge might hold for the human engineer. Therefore, the human engineer classifies

movements in other ways; movements toward and away from the body; blind positioning movements vs. movements with visual feedback; actual movements vs. "static movements"; repetitive vs. non-repetitive movements; rotary vs. linear movements, etc.

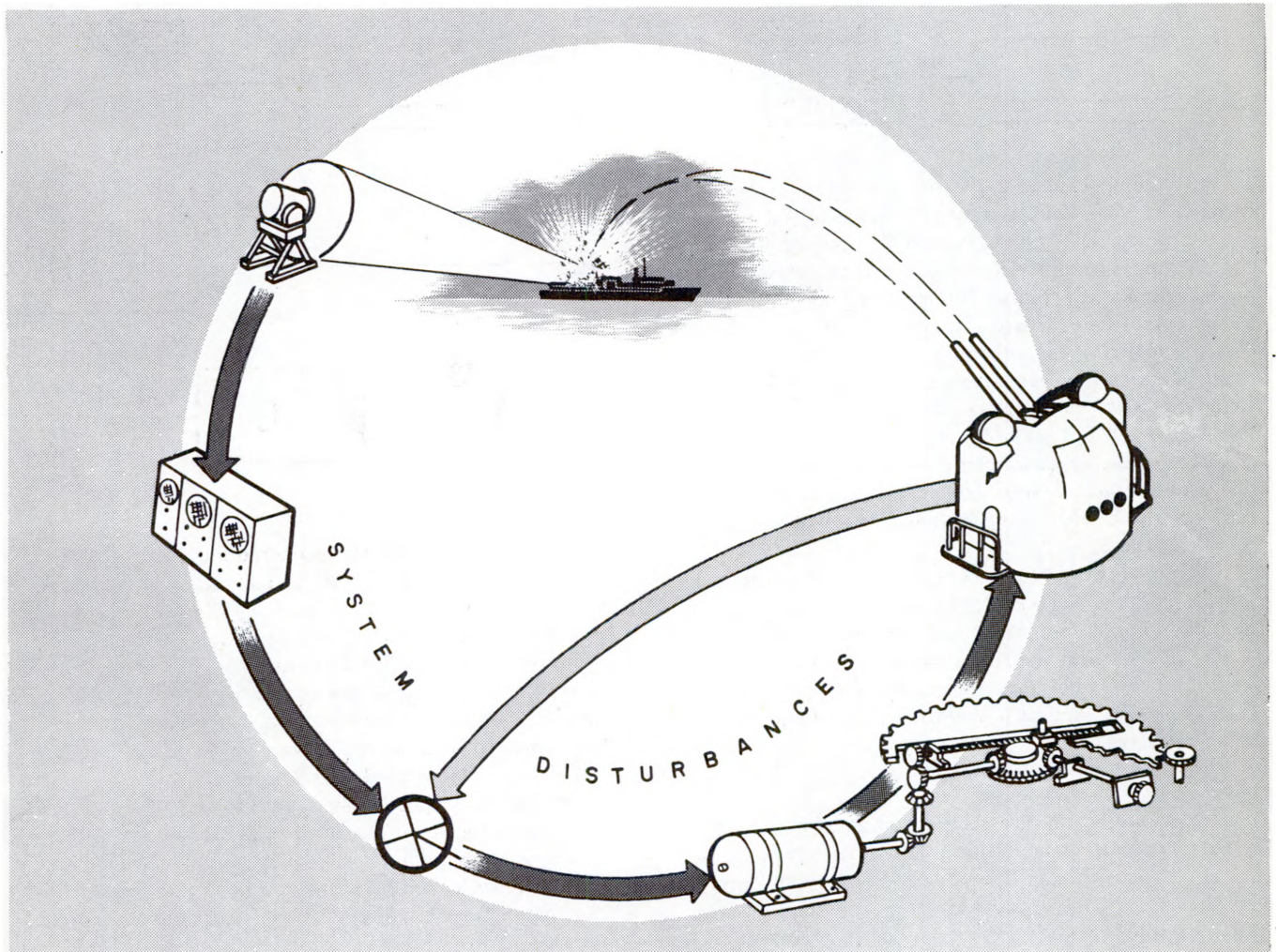
For example, the human engineer studies static movements, i.e., those movements requiring the holding of something steady without tremor. The results of these studies reveal that there is less tremor if some friction must be overcome, if the operator is relaxed, and if he is not fatigued. The studies also reveal there is less tremor when seated than when standing, and less tremor when the arm is supported rather than when it is free. In the study of repetitive motion, for example, the results have aided the human engineer in the design of better telegraph keys. These results show the average rate of tapping ranges between 8 and 13 times per second. Also, for a given individual, the best tapping performance is obtained horizontally rather than vertically and with the wrist or the whole arm rather than the finger.

A most important type of movement studied by the human engineer is tracking. In this motion, continuous adjustments are made at relatively low speed to stay on a continuously moving target. This study revealed the undesirability of friction in tracking and the desirability of inertia in the tracking mechanism.

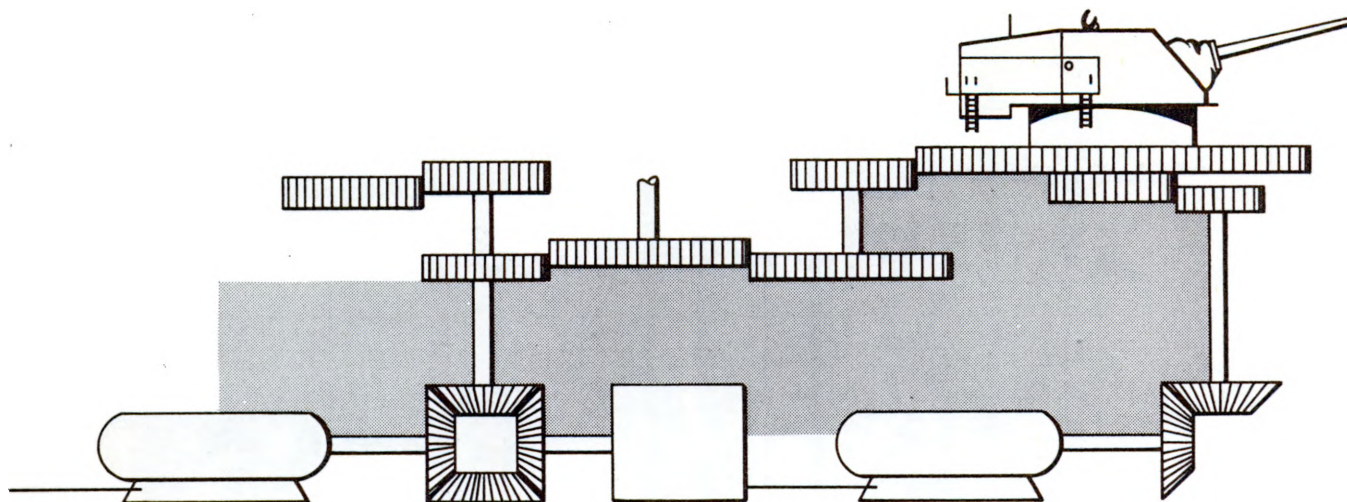
Remaining for the human engineer to determine are fatigue factors, layout, and time and motion study.



Introduction to SYSTEM DYNAMICS



System dynamics is the study of the problem of controlling the motion of the various components of a weapons system. The methods and terms used in the analysis of the dynamics of systems are discussed first. The basic problems which must be solved by the control system are then analyzed. Finally, an introduction to basic control system configurations used in weapons systems is provided.



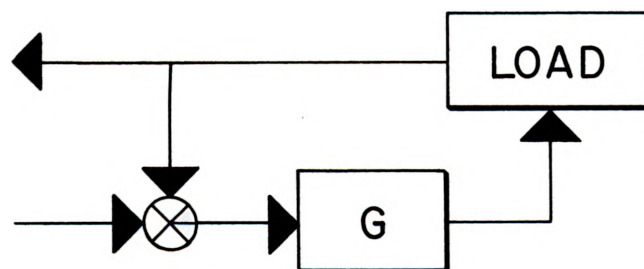
The basic component of a control system is the servo-mechanism. The fundamental principles of this component of a weapon system were discussed in Volume 1. This chapter covers the dynamics of the system only; the components are treated as "black boxes" for which the input and output functions are known. The expression which gives the output of the box as a function of the input signal is called the "transfer function" of the device. Various aspects of transfer functions are discussed in section 1.

A number of factors contribute to possible error in the motion of weapons system components. The resultant error creates problems which must be solved by the control system.

The three major sources of error are tracking noise, dynamic lag, and weapon station motion.

Tracking noise is any disturbance in the operation of the tracking sensor, usually a radar set. This noise arises from any one of three possible sources, each of which is predominant at a specific range of operation of the radar set.

- 1) Internal noise, which is important at long ranges
 - 2) Pulse-to-pulse modulation by the target, which is important at all ranges (in conical-scan radars only)
 - 3) Wandering of the apparent center of reflection on a large, complex target (very important at close ranges).
- Increased dish size and radiated power of modern radar, coupled with receiver improvements, has greatly increased usable tracking ranges and thereby reduced the importance of internally generated noise at firing ranges where accurate tracking is necessary. Pulse-to-pulse modulation noise, which is target generated and almost independent of range, cannot be reduced in conical-scan radars. Although monopulse radars are not affected by pulse-to-pulse modulation, they offer improvement only at medium ranges (greater than 1500 yards) where the effect predominates. Wander noise affects both conical-scan and monopulse radars and is predominant at close ranges (less than 1500 yards) where the target dimensions become too large to achieve the desired tracking accuracy.



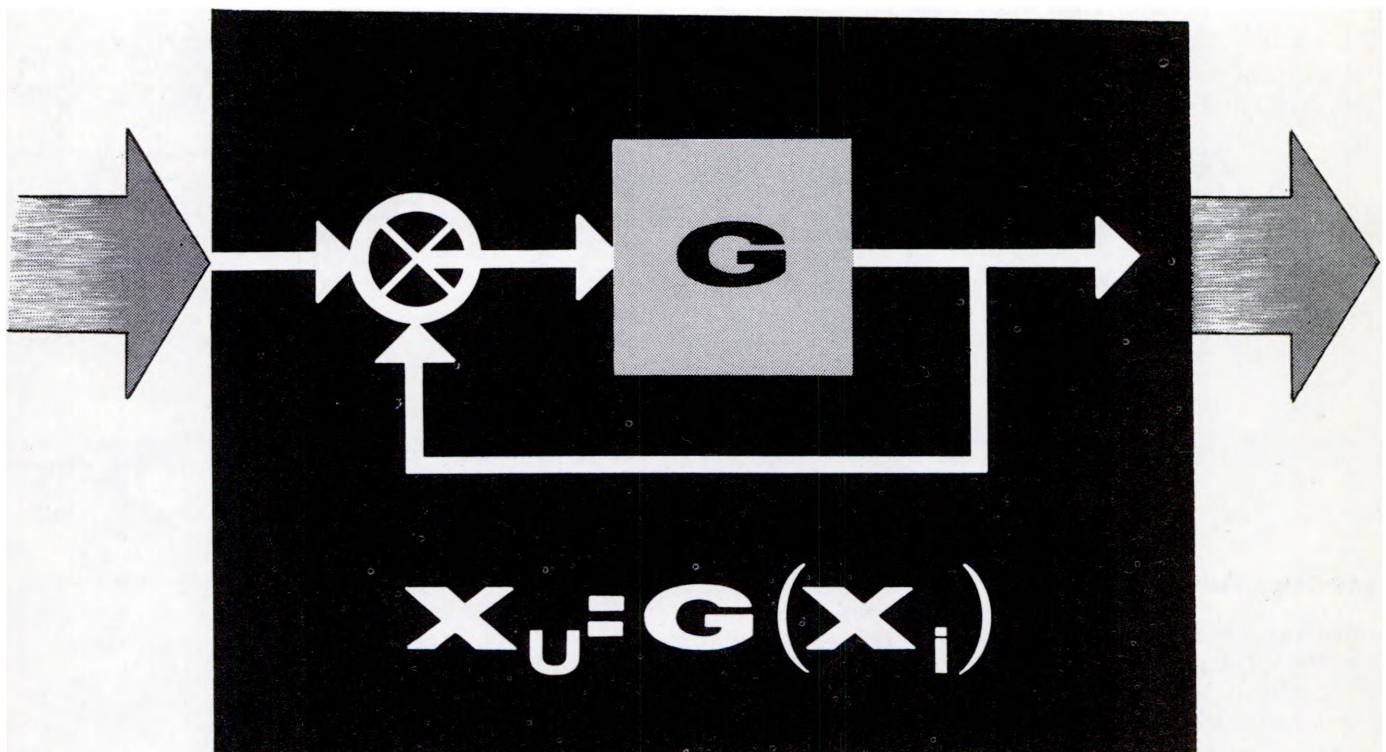
Thus, irreducible target-generated angular tracking noise really represents the fundamental limitation on tracking accuracy.

While the system is following the target it must be effective in offsetting disturbances so that the errors caused by these disturbances can be kept small. For ground-based installations there may be a number of disturbing torques on the antenna and weapon drives which result from wind, friction, etc.

Airborne and shipboard systems have in addition the problem of motion of the weapon station, the platform on which the tracking sensor and weapon are mounted. Disturbances resulting from weapon-station motion may be a much more serious problem than disturbing load torques. While following a target a weapon-control system tends to oppose the effect of weapon-station motion. However, tracking errors can be quite large if there is much weapon-station motion. Consequently, to counteract weapon-station motion, gyros are often employed to measure it independently of target motion. The gyro signals, which are essentially noise free, are used to provide space stabilization of the tracking and weapon line. These three factors affecting the dynamic stability of the system are discussed in detail in section 2.

Weapon control systems may be classified in either of two distinct manners: 1) by the presence or absence of a feedback loop; 2) by which component is driven independently. The various control system configurations used in naval weapons systems are discussed in section 3.

DEFINITION OF TERMS



This section defines the basic terms applied to feedback control system variables and the various criteria used to analyze these systems. The first topic is the symbology employed for describing the action of the basic building blocks of weapon control systems: the transfer function. The second topic is a discussion of methods used to describe the variations in system response with different input frequencies. Finally, the various orders of control systems are defined and discussed.

TRANSFER FUNCTIONS

definition

An essential tool for a mathematical treatment of a process is an expression for the output variable of the process in terms of the input variable. The starting point is an expression relating the input and output variables for a single element. Depicting the element as a block, the factor G represents the transfer function of the element and of the open loop system.

$$\text{output} = G (\text{input})$$

The transfer function is an operator, i.e., the function is not necessarily multiplied by the input variable. If the element is a simple amplifier which increases the input variable by a factor A , then the transfer function would simply be multiplied by the input function.

$$Y = G(X)$$

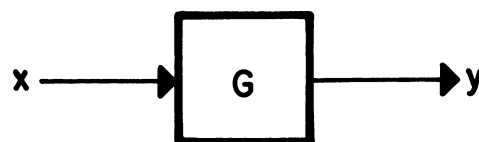
$$G = A$$

$$y = Ax$$

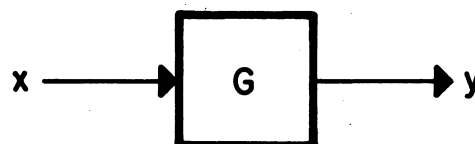
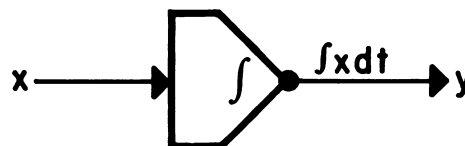
If the element is an integrator, however, the transfer function would not act as a multiplier.

$$y = \int x \, dt$$

$$G = \int dt$$



$$\frac{\text{OUTPUT}}{\text{INPUT}} = \frac{Y}{X} = G$$



$$G = \int dt$$

application

One source of problems in the design of dynamic systems is the lack of perfect elements. The transfer function of an element will generally contain a time lag and may depart from the design value due to construction tolerances. Therefore, it is often important to minimize the effect of a deviation (ΔG) from the ideal transfer function (G). The effect of a deviation in the simple open-loop system can be analyzed as follows:

$$\begin{aligned} y &= Gx \\ y + \Delta y &= (G + \Delta G)x \\ &= Gx + \Delta Gx \\ &= y + \Delta Gx \\ \Delta y &= \Delta Gx \end{aligned}$$

The change in output (Δy) is directly proportional to the change in the element (ΔG).

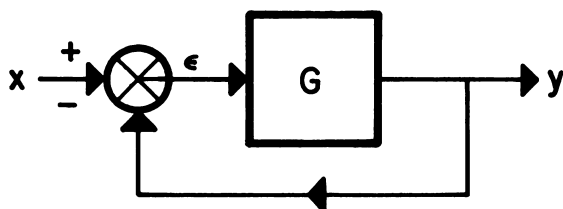
Applying the same procedure to the simple closed-loop system:

$$\begin{aligned} y &= \frac{G}{1 + G} x \\ y + \Delta y &= \frac{G + \Delta G}{1 + G + \Delta G} x \\ &= \frac{G}{1 + G + \Delta G} x + \frac{\Delta G}{1 + G + \Delta G} x \end{aligned}$$

If ΔG is small compared with 1,

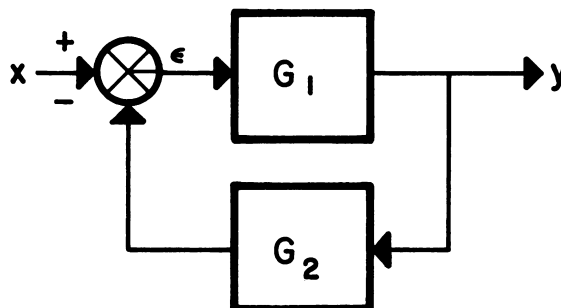
$$\begin{aligned} 1 + G + \Delta G &\approx 1 + G \\ y + \Delta y &= y + \frac{\Delta G}{1 + G} x \\ \Delta y &= \frac{\Delta G}{1 + G} x \end{aligned}$$

The transfer function for the single open-loop element is trivial. The value of the method lies in the inclusion of more complex arrangements, the first of which is the single-element closed-loop system. As derived in the illustration, the transfer function of the system is $G/(1+G)$, where G is the transfer function of the element.



$$\begin{aligned} y &= G \epsilon \\ \epsilon &= x - y \\ y &= G(x - y) \\ &= Gx - Gy \\ y(1 + G) &= Gx \\ \frac{y}{x} &= \frac{G}{1 + G} \end{aligned}$$

A slightly more complex system is the two-element closed-loop system illustrated. The transfer function of the system is $G_1/(1 + G_1 G_2)$, where G_1 is the transfer function of the in-line element and G_2 is the transfer function of the feedback element.



$$\begin{aligned} y &= G_1 \epsilon \\ \epsilon &= x - G_2 y \\ y &= G_1(x - G_2 y) \\ &= G_1 x - G_1 G_2 y \\ y(1 + G_1 G_2) &= G_1 x \\ \frac{y}{x} &= \frac{G_1}{1 + G_1 G_2} \end{aligned}$$

The error is reduced by the factor $1/(1 + G)$ compared with the error for the open-loop system. If the element transfer function is large, the error can be greatly reduced by the use of the closed-loop system. If G is small, the error can be reduced by the use of the two-element closed-loop system. The same analysis of this system yields:

$$\begin{aligned} y &= \frac{G_1}{1 + G_1 G_2} x \\ y + \Delta y &= \frac{G_1 + \Delta G_1}{1 + G_1 G_2 + \Delta G_1 G_2} x \\ &= \frac{G_1}{1 + G_1 G_2 + \Delta G_1 G_2} x + \frac{\Delta G_1}{1 + G_1 G_2 + \Delta G_1 G_2} x \end{aligned}$$

Ignoring the $\Delta G_1 G_2$ term,

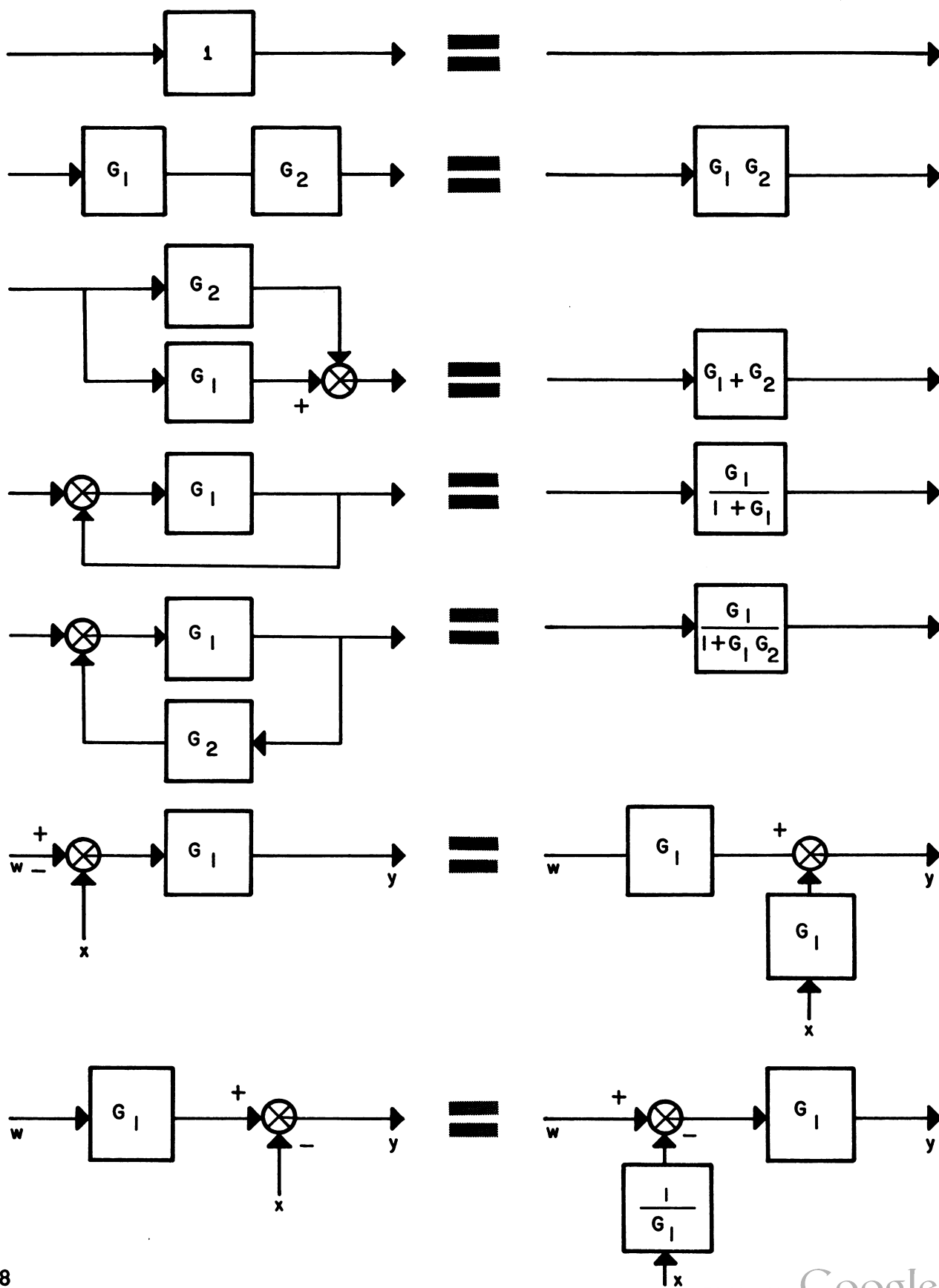
$$\begin{aligned} y + \Delta y &= y + \frac{\Delta G_1}{1 + G_1 G_2} x \\ \Delta y &= \frac{\Delta G_1}{1 + G_1 G_2} x \end{aligned}$$

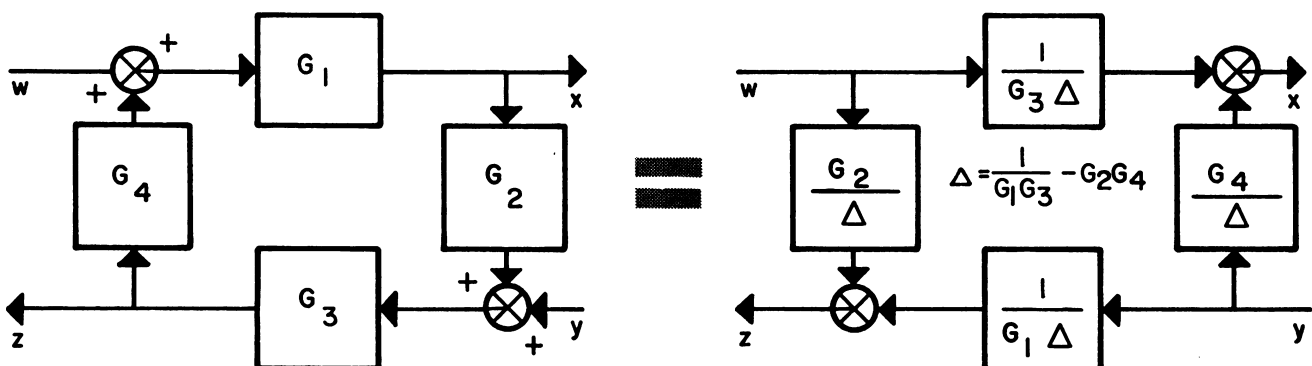
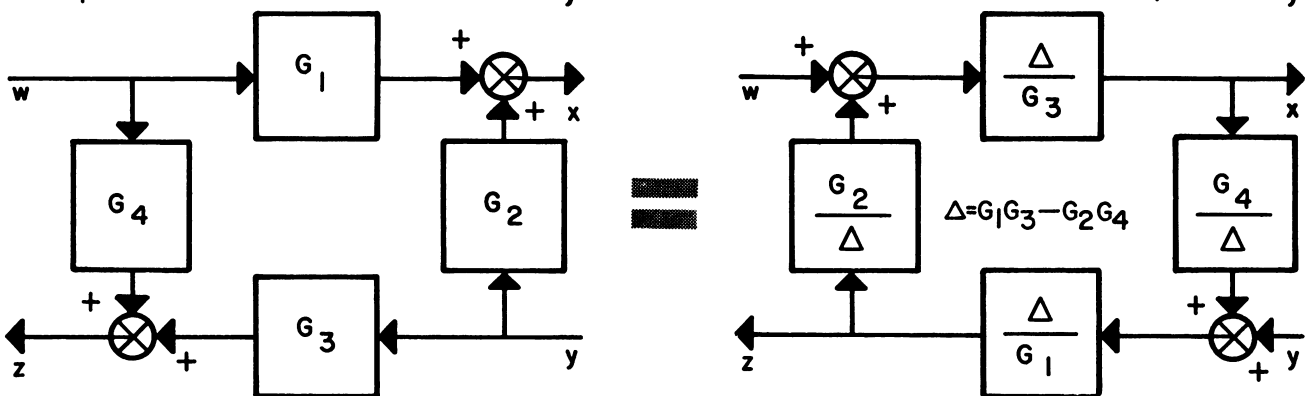
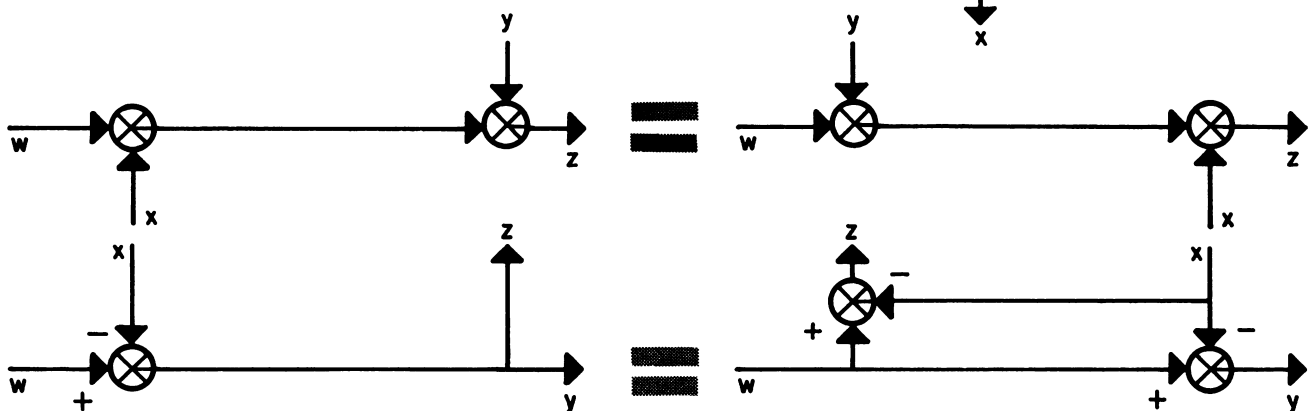
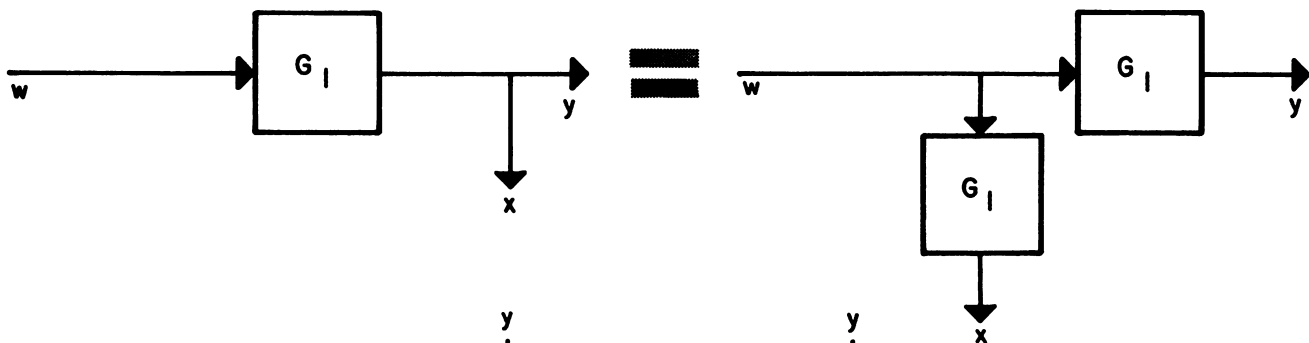
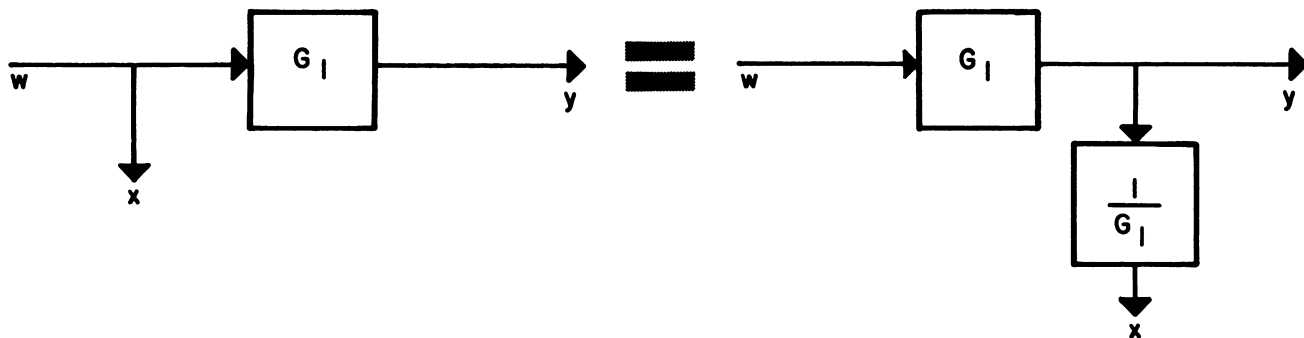
The error is small if G_2 is large."

reduction

Since the end product of the system is the important parameter to the analyst, a small system can be replaced by a single block with a transfer function equivalent to that of the system replaced. The two closed-

loop systems just discussed, plus several other simple combinations, are illustrated along with their equivalent diagrams. These equivalence rules can be used to manipulate complex diagrams.





BASIC PARAMETERS

To analyze the performance of a multiloop fire control system, it is necessary to break down the complex response of the system into the simpler actions of its individual loops. To perform this breakdown, a precise terminology is required for defining the elements and characteristics of a feedback-control loop. The following section presents the required terminology and describes some of the basic feedback-control loop characteristics.

For explanatory purposes, a typical signal-flow diagram for a simple feedback loop is given. The variables shown are:

$$\begin{aligned} X_i &= \text{loop input} \\ X_e &= \text{loop error} \\ X_u &= \text{loop return signal} \end{aligned}$$

The transfer function connecting X_e and X_u is designated G , and is called the loop transfer function:

$$G = \frac{X_u}{X_e} = \text{loop transfer function}^1$$

Two other transfer functions relating the loop variables are designated with double subscripts as follows:

$$G_{ie} = \frac{X_e}{X_i} = \text{error transfer function}$$

$$G_{iu} = \frac{X_u}{X_i} = \text{return transfer function}^2$$

In general, $G_{ab} = \frac{X_b}{X_a}$.

According to the double-subscript notation, the loop transfer function, G , could also be designated as G_{eu} , but because this loop transfer function is so frequently used, the simple designation G is more convenient.

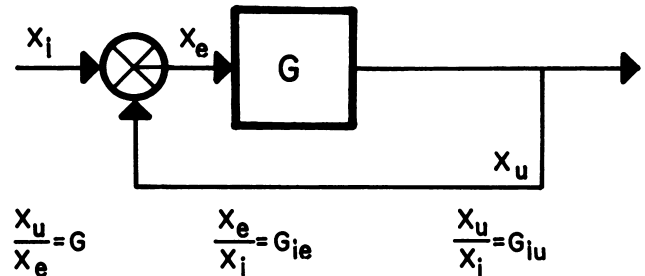
The relation among the three transfer functions of the feedback control loop is:

$$\begin{aligned} G_{iu} &= \frac{G}{1+G} \\ G_{ie} &= \frac{1}{1+G} \end{aligned}$$

notes

¹ The function G is often called the open-loop transfer function.

² The return transfer function is often called the closed-loop transfer function.



Parenthetical subscripts are used throughout the following paragraphs to designate the transfer functions of the various loops within a fire-control system. The loop transfer functions of the primary feedback-control loops are designated as follows:

tracking loop: $G(t)$

weapon loop: $G(w)$

stabilization loop: $G(s)$

antenna loop: $G(a)$

The same identifying subscripts are used to designate the return and error transfer functions for these loops. For example, the return and error transfer functions of the tracking loop are designated

$G_{iu}(t)$ = tracking-loop return transfer function

$G_{ie}(t)$ = tracking-loop error transfer function

The transfer functions for the feedback-control loops designated above are idealized in that they do not, in general, define the complete transfer functions of these control loops under actual operation in a multiloop system.

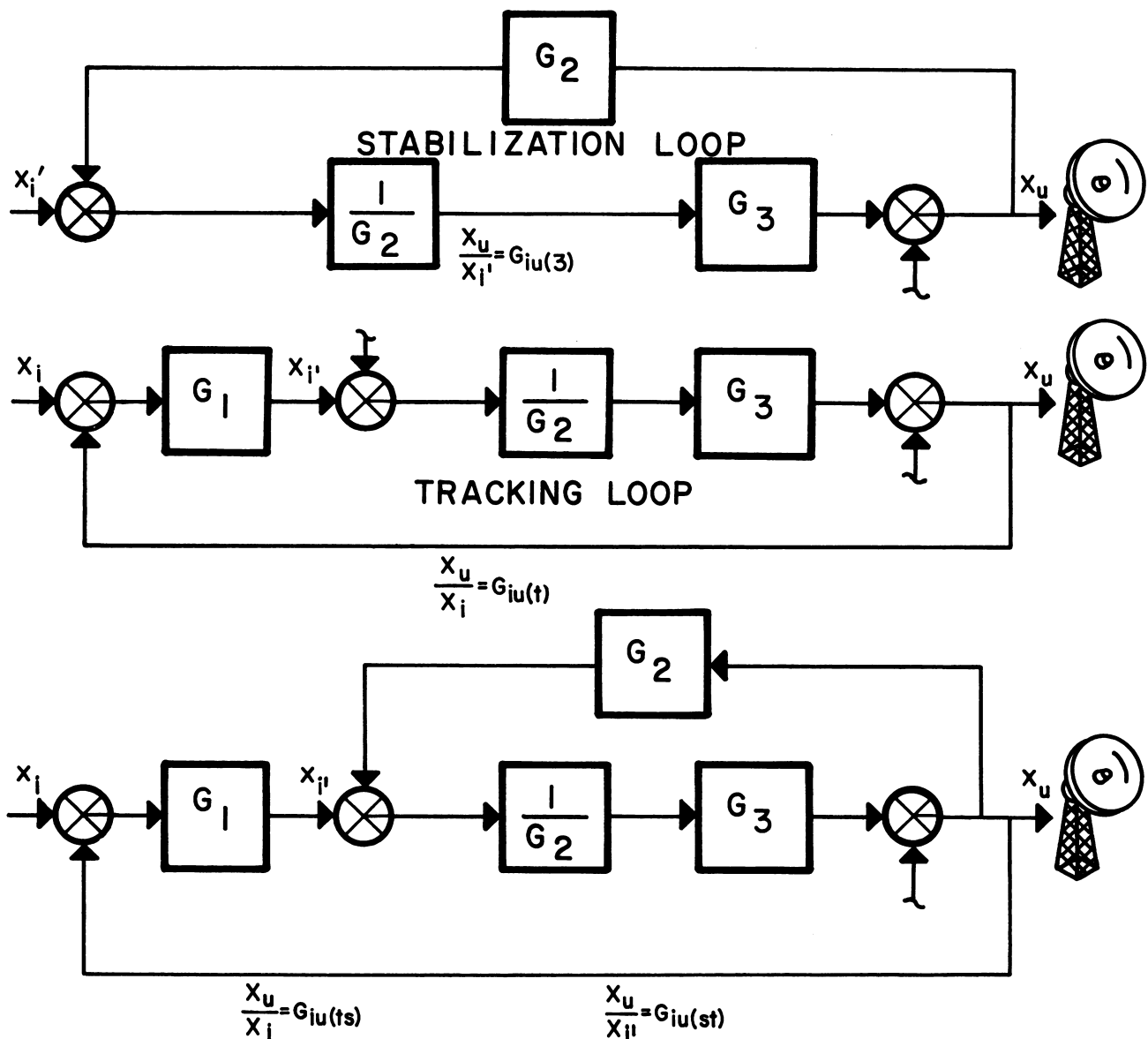
For the purpose of analysis, it is desirable to think of the system as being basically composed of independent, primary loops, but remembering that in actual operation these loops affect one another so that the primary loops are changed to compound loops. To designate the transfer function for a compound loop, multiple parenthetical subscripts are used. The first letter in the subscript indicates the primary loop, and the second letter the secondary loop.

OF FEEDBACK LOOPS

For example, in a stabilized antenna, the transfer function $G_{(s)}$ is defined as the loop transfer function of the stabilization loop with the tracking loop opened. When the tracking loop is closed, the tracking loop affects the loop transfer function of the stabilization loop and the resultant stabilization-loop transfer-function is $G_{(st)}$, designated with a double parenthetical subscript. The function $G_{(st)}$ is called the loop transfer function of the stabilization loop with the effect of the tracking loop included. The stabilization loop also affects the tracking loop; and the resultant loop transfer function of the tracking loop with the effect of the stabilization loop included is designated $G_{(ts)}$.

In analyzing a compound loop, it is important to recognize the corresponding primary loop, because a compound loop has the same asymptote-crossover frequency (and hence approximately the same gain-crossover frequency) as its primary loop. Thus, the function $G_{(st)}$ of

the compound stabilization loop has the same asymptote-crossover frequency as the function $G_{(s)}$ of the primary stabilization loop. Similarly, $G_{(ts)}$ for the tracking loop has the same asymptote-crossover frequency as $G_{(t)}$. It might appear in abstract that the above symbolism employed to designate the various loops is rather arbitrary and excessively complex. The loop transfer functions could have been designated in a general fashion as G_1, G_2, G_3 , etc., but this would not have emphasized the basic loop actions. Despite the complexity of a multi-loop control system, the individual loops must operate essentially independently of one another in order for the system to be adequately stable. Consequently, it is necessary to examine the individual loops as separate entities, in order to study the performance of the complete system. The reason for the symbolism is to help show the manner in which the over-all system response is built up from the actions of the individual loops.



GAIN-FREQUENCY

In general, the magnitude of G will vary with the frequency of the input variable. In other words, the response of the system will be different for input signals of different frequencies.

The magnitude of the loop transfer function, for real frequencies, is called the loop gain:

$$G(j\omega) = \text{loop gain}$$

The term loop gain is also frequently used to define a constant representing the zero-frequency value of G , sG or s^2G , depending upon whether the loop has 0, 1, or 2 integrations at zero frequency.

logarithmic frequency-response plots

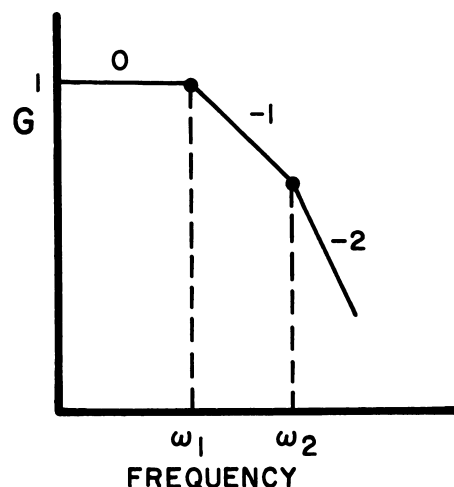
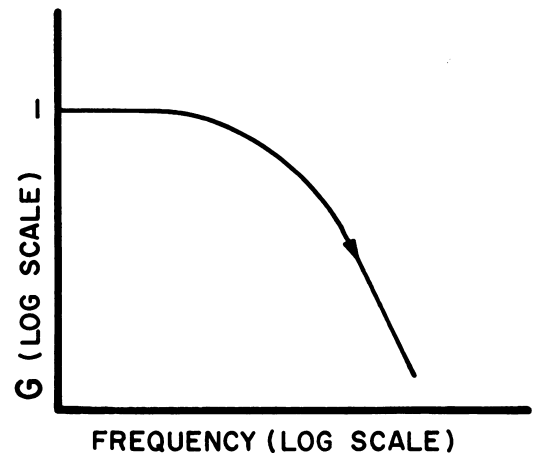
The frequency-response plots in the following paragraphs show magnitude versus frequency on logarithmic axes. In many cases asymptotes are used for the magnitude curve, instead of the curve itself.

The manner in which the frequency-response plots are presented is illustrated. It is an asymptotic magnitude plot of the frequency response of the function

$$G = \frac{1}{[\frac{s}{\omega_1} + 1][\frac{s}{\omega_2} + 1]}$$

The labels "log scale" are shown on the two axes of the graph to indicate that they are logarithmic. These labels are normally omitted in frequency-response plots, and although all of these plots have logarithmic scales, the labels on the scales represent linear rather than logarithmic units. The label 1 on the vertical scale of the graph indicates that the low-frequency value of G is unity. The values ω_1 and ω_2 on the horizontal scale represent frequency in radians per second. The slopes of the magnitude asymptotes in the graph are indicated by the numbers 0, -1, and -2. Each number represents the logarithmic slope of the corresponding asymptote. The logarithmic slope of the magnitude of a frequency response represents the derivative of the logarithm of the magnitude with respect to the logarithm of the frequency.

Thus, a logarithmic slope of -1 indicates that the asymptote is proportional to $1/\omega$ while a slope of -2 indicates that the asymptote is proportional to $1/\omega^2$. This means of identifying slope is more convenient than such terms as -6 decibels per octave, -20 decibels per decade, or -10 decibels per decade; all of which designate a slope of -1 as defined here.



RELATIONSHIPS

gain-crossover frequency

Frequency response of a control system depends upon the range of frequencies over which the input order may oscillate and still produce acceptably similar oscillations in the output.

For example, a ship in which the rudder is moved slowly back and forth (low frequency) will follow the position of the rudder in the desired manner. However, when the rudder is moved very rapidly back and forth (high frequency), the ship does not respond to positions of the rudder and continues on a straight course. In general, there is some upper limiting frequency of rudder motion (oscillation) beyond which the ship will not respond. The range of frequencies within which the ship's-rudder system does respond in the designed fashion is called the bandwidth of the system. The bandwidth of any system is the range of frequencies within which the performance of the system, with respect to some characteristic, falls within specified limits. In general, the bandwidth of a control system is the frequency band between zero and some upper limiting frequency. A very important characteristic of the transfer function of a feedback control loop is its gain-crossover frequency, defined as the frequency at which the loop gain passes through unity with negative slope*, and is designated ω_{cg} . It is the most reliable parameter for describing the bandwidth of a system. Other measures of bandwidth based on the closed-loop magnitude ratio, such as the frequencies for peak magnitude ratio, unity magnitude ratio, or the -3db magnitude ratio, are not closely related to system performance.

An important advantage of using the gain-crossover frequency as a measure of bandwidth is that it determines quite reliably some important characteristics of the transient responses. The time for the step response to rise to 63 percent of its final value, which is called the rise time, is roughly equal to $1/\omega_{cg}$. If the slope of the loop gain is reasonably steep at frequencies below gain crossover, the maximum error to a unit ramp is roughly $1/\omega_{cg}$; if the slope of loop gain is reasonably steep at frequencies above gain crossover, the maximum value of the return response to a unit impulse is roughly ω_{cg} .

note

* If a feedback control loop has a loop gain less than unity at zero frequency, there are two frequencies where the loop gain is unity: one at low frequency where the slope of G is positive and one at high frequency where the slope of G is negative. It is the higher of these two frequencies that represents the gain-crossover frequency.

asymptote-crossover frequency

For calculation purposes it is usually more convenient to determine the frequency at which the asymptote of the loop gain, rather than the actual loop gain, crosses unity. Consequently, it is desirable to define an asymptote-crossover frequency that can be used as an approximation to the actual gain-crossover frequency. In the analysis presented in later pages, the asymptote crossover frequency is referred to as simply the crossover frequency because the analysis is not precise enough to take account of the small difference between the gain-crossover frequency and the asymptote-crossover frequency.

For adequate stability, the phase lag of a feedback-control loop should not greatly exceed 135 degrees at gain

crossover. Consequently, the logarithmic slope of the loop gain should not be much steeper than -1 at gain crossover. Because of this limitation upon the slope of the loop gain, the loop transfer functions of almost all practical feedback-control loops approach an asymptote with a logarithmic slope of -1 in a frequency region at gain crossover or quite close to gain crossover. The frequency where this asymptote (extended if necessary) crosses unity gain is defined as the asymptote-crossover frequency. This frequency variable is very convenient because it can be related easily to the gains built within the various elements of a feedback-control loop.

The definition of the asymptote crossover frequency is illustrated by two graphs, (a) and (b). In plot (b), the asymptote of G has a slope of -2 at gain crossover (ω_{cg}), but the asymptote crossover frequency ω_c is the frequency where the extension of the asymptote of -1 slope crosses unity gain. In plot (a), there are two frequency regions where the asymptote slope is -1 , one at very low frequencies and the other at gain crossover. As the plot shows, it is the asymptote near gain crossover that defines ω_c . The intersection of the low-frequency asymptote with the unity-gain axis represents the velocity constant K_v .

The asymptote crossover frequency is defined to make the following rules generally hold:

- 1) The asymptote crossover frequency ω_c is fairly close to the actual gain crossover frequency ω_{cg} .
- 2) The loop transfer function approaches the asymptote ω_c/s in a frequency region at gain crossover or close to it.

Because of the second property, the asymptote crossover frequency can be conveniently used as the gain factor in the expression for the loop transfer function.

To illustrate the use of the asymptote crossover frequency ω_c as the gain factor for a feedback-control loop, consider the loop transfer function illustrated in plot (a). In the region near ω_{cg} , the loop transfer function G approximates the asymptote:

$$G \approx \frac{\omega_c}{s} \quad \text{for } \omega \approx \omega_{cg}$$

The high frequency factors, with break frequencies at ω_3 and ω_4 , can be added to this expression without changing the asymptote near gain crossover, by expressing these factors in the form $[(s/\omega_x) + 1]$. Thus G can be approximated as follows in the mid- and high-frequency region:

$$G \approx \frac{\omega_c}{s} \left[\frac{1}{\left[\frac{s}{\omega_3} + 1 \right] \left[\frac{s}{\omega_4} + 1 \right]} \right] \quad \omega \geq \omega_{cg}$$

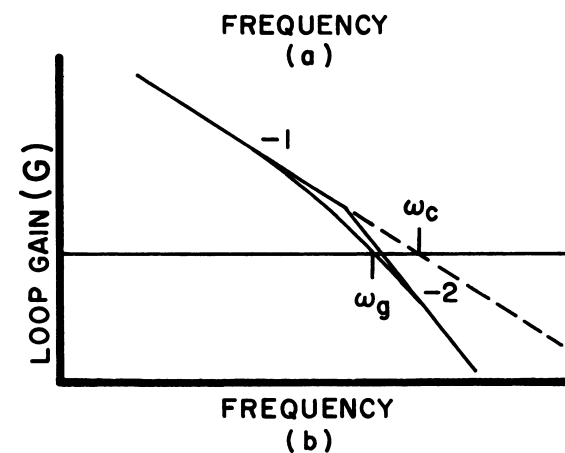
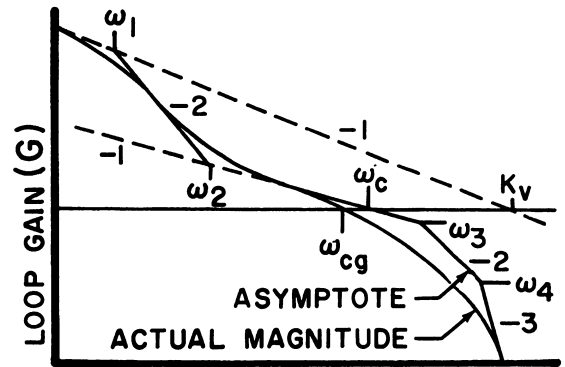
By expressing the low-frequency factors, with break frequencies at ω_1 and ω_2 , in the form $[1 + (\omega_x/s)]$, they can be added to this expression without changing the asymptotes near gain crossover and at higher frequencies. Including the effect of these low-frequency factors, the complete expression of G is

$$G = \left[\frac{\omega_c}{s} \right] \left[\frac{1 + \frac{\omega_2}{s}}{1 + \frac{\omega_1}{s}} \right] \left[\frac{1}{\left[1 + \frac{s}{\omega_3} \right] \left[1 + \frac{s}{\omega_4} \right]} \right]$$

Low-Frequency Factors High-Frequency Factors

Multiplying the factors out gives

$$G = \left[\frac{\omega_c}{s} \right] \left[\frac{s + \omega_2}{s + \omega_1} \right] \left[\frac{\omega_3 \omega_4}{(s + \omega_3)(s + \omega_4)} \right]$$



The two expressions for G are complete and exact, because all the dynamic factors are included.

The above technique for writing the loop transfer function expresses the gain factor in terms of the asymptote of -1 slope near gain crossover. In contrast, the usual technique for writing this function is to express the gain factor in terms of the low-frequency asymptote. For example, the loop transfer function, illustrated previously as plot (a), approaches the following asymptote at low frequencies:

$$G \approx \frac{K_v}{s} \quad \text{for } \omega \approx 0$$

where K_v is the velocity constant. The break frequencies of all the factors are higher than the frequency range of this asymptote; for them to approximate unity at low frequencies, all of the factors should be expressed in the form $[1 + (s/\omega_x)]$. Thus, the complete expression for G is:

$$G = \frac{K_v}{s} \left[\frac{1 + \frac{s}{\omega_1}}{\left[1 + \frac{s}{\omega_2} \right] \left[1 + \frac{s}{\omega_3} \right] \left[1 + \frac{s}{\omega_4} \right]} \right]$$

This expression is equivalent to the one derived previously except that the gain factor is the velocity constant K_v , rather than the asymptote crossover frequency ω_c . The relation between these two gain factors, K_v and ω_c , can be determined by setting the last equation equal to the one derived previously which gives:

$$K_v = \frac{\omega_2}{\omega_1} \omega_c$$

ORDER OF CONTROL

In order to gain the most favorable response, various combinations of zero, first, and second order control may be employed in a system. Order of control is an indication of the relationship between the input to the control system and the output of the control system. By the positioning of an automobile accelerator, first order control is achieved. For the position input of the accelerator, you are getting as an output a constant vehicular velocity or the first derivative of the input. Thus, when we state that control is of a first order, or second order, we are merely stating that for a position input we are achieving as an output, velocity or acceleration, respectively. To better understand these orders of control systems, we will examine them as applied to a tracking loop.

zero order (position control)

The simplest type of control is zero order or position control.

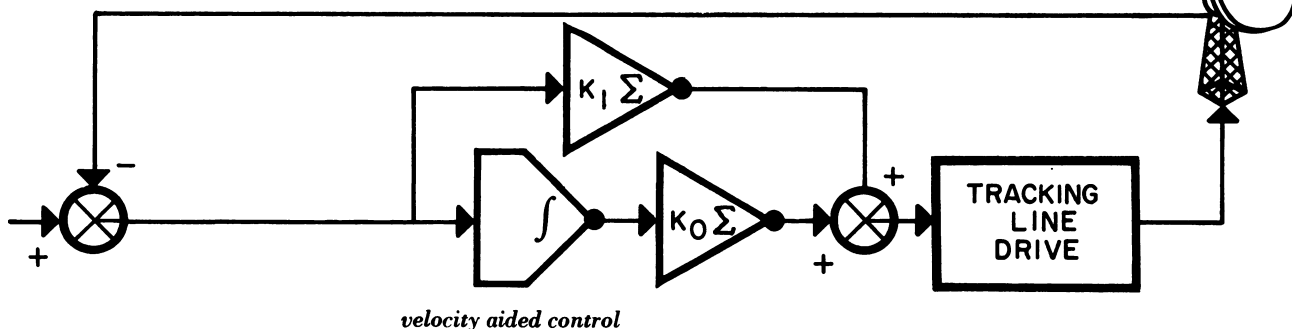
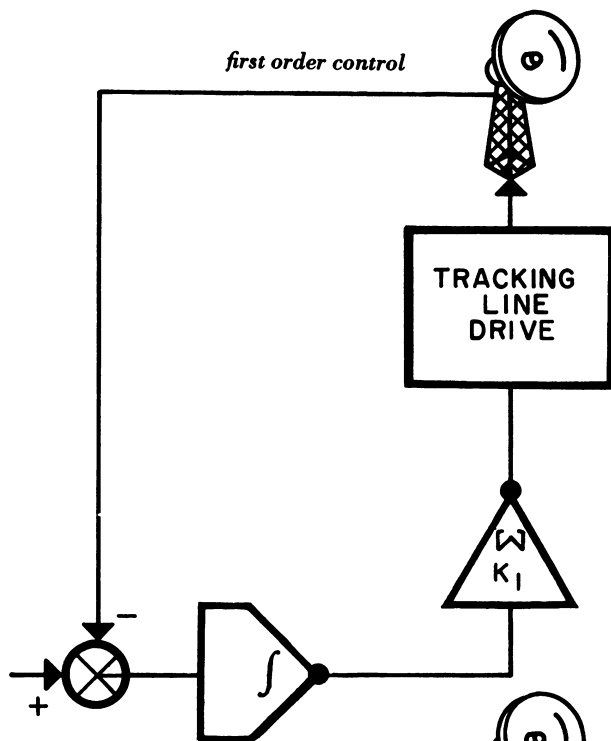
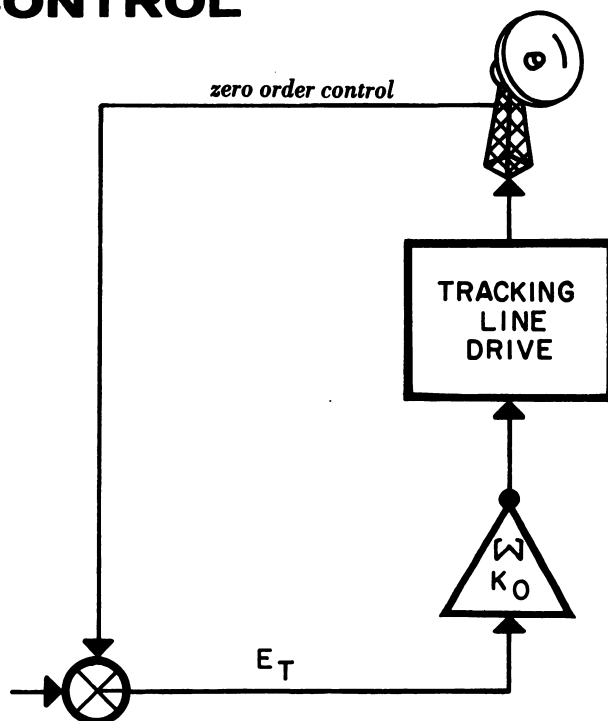
The value of K_0 is a constant which represents the gain, amplification, or gearing ratio, in the control. A position input of the tracking line error (E_T) will produce a control system position output which opposes the error. When there is no error there is no output.

first order (velocity control)

The next order in control system complexity, first order or velocity control, is one in which the tracking line error, position input, causes an output signal which is changing at a constant rate. It should be noted that as long as a positive input to the control system exists, the control system output will continue to increase in magnitude. When the error is reduced to zero, the system will continue to produce an output.

A variation of the first order control, which combines both zero order and first order control, is velocity-aided control. In the design of a velocity-aided system, an extremely important consideration is the proper selection of values for the constants, K_0 (position) and K_1 (velocity). By varying the velocity-aiding constant (the ratio K_0/K_1), control system characteristics can be modified so the output reflects a large amount of velocity-aided control, or primarily position control with a small change to the velocity output of the system for any given input.

The best all around velocity constant appears to be about 0.5, with values between 0.2 and 0.8 constituting an acceptable range.



Velocity-aided control in a tracking system is often referred to as aided tracking. Investigation has shown that aided tracking, in general, gives more satisfactory results than the use of either position or velocity control alone. Why this should be so may be seen from the following:

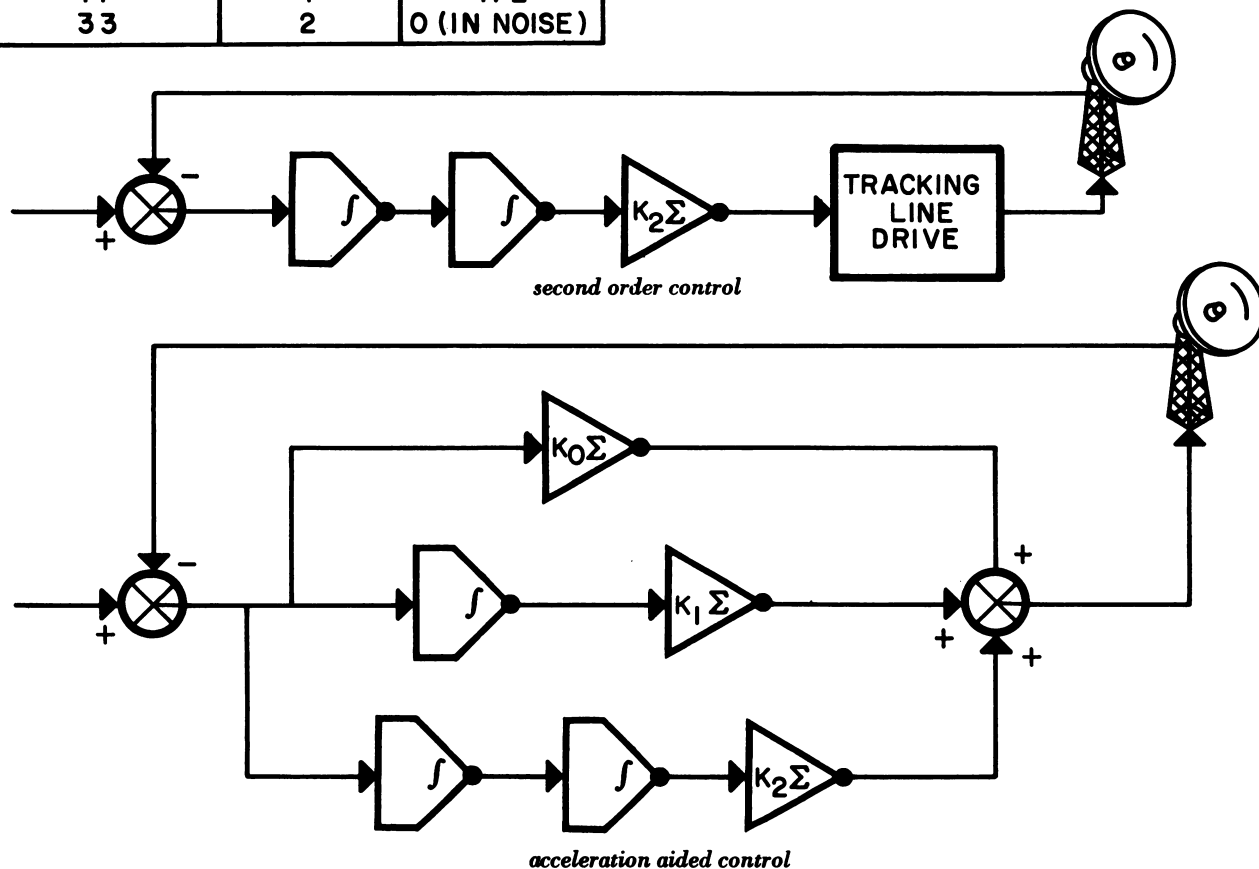
- 1) A tracking system, using aided tracking, can track a target whose angular velocity is constant merely by keeping the control system input fixed, while with zero order control tracking the tracking line error input to the control system must be constantly repositioned.
- 2) For slowly changing target velocity, the tracking system can correct for any angular distance it has fallen behind by putting in an additional displacement of the control system position input (E_t). This has the effect of simultaneously changing the position of the tracking line and increasing its angular velocity as well. When the tracking line again falls behind the target, all that is needed is another slight increment to the position input.
- 3) Aided tracking helps the system continue tracking through a region in which the target is temporarily not visible.
- 4) Experience shows that aided tracking is more stable than velocity tracking, in that there is less tendency for the tracking system to hunt about the line of sight or lag as illustrated.

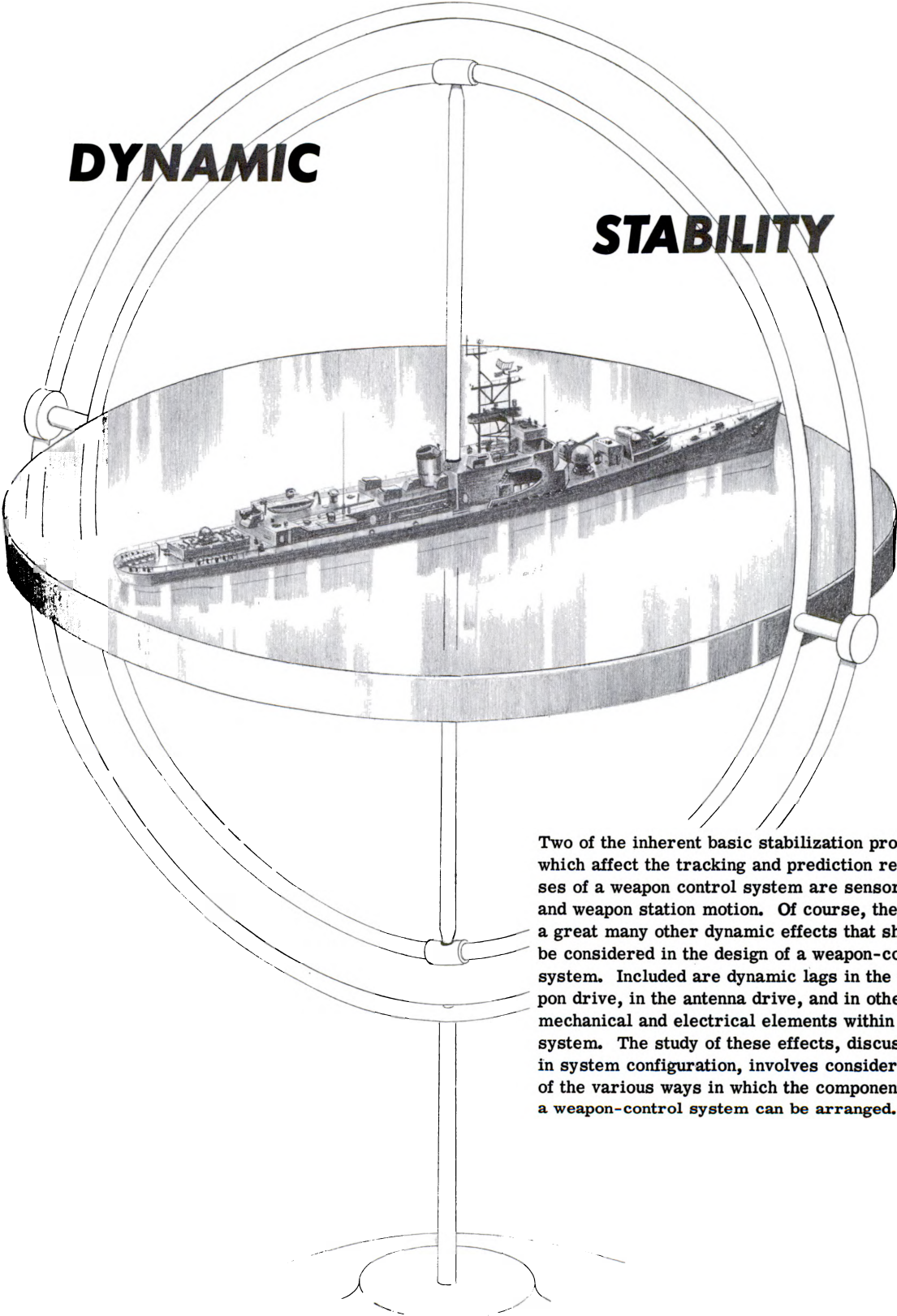
TRACKING LOOP GAIN	TRACKING LAG	
	UNAIDED	AIDED
8.5	8	1
17	4	1/2
33	2	0 (IN NOISE)

second order (acceleration control)

The next step in complexity of control is second order or acceleration control. Upon the introduction of a position input to the control system, the output rate will increase or the velocity of the output will increase. A second order system may often include the two lower order terms, position and rate control; such a system is referred to as an acceleration-aided system. The selection of the constants for this type of control system is rather critical and the ratio of position to velocity to acceleration ($K_0:K_1:K_2$) should be taken to have values in the range 8:4:1 or 8:2:1. Thus the control system would heavily weight the position output, next give a smaller correction to the velocity of the output, and finally, modify the acceleration characteristic of the output by only a small amount.

Although control systems of still higher order may occur, they follow the general patterns illustrated. The number of terms used in aiding should equal or exceed by one the derivative of the input which is constant. For example, in tracking a constant velocity target the first derivative of position (velocity) would be constant, and the second derivative (acceleration) would be zero. A second or third order control system is indicated. Control systems of higher order are rare. One example is the depth control of a submarine, which is a fourth order system. The planesman's wheel controls the rate of turn of the bow planes. The plane angle, in turn, controls the rate of pitch of the boat, and the pitch angle controls the rate of descent. Depth, therefore, is the fourth order output.





Two of the inherent basic stabilization problems which affect the tracking and prediction responses of a weapon control system are sensor noise and weapon station motion. Of course, there are a great many other dynamic effects that should be considered in the design of a weapon-control system. Included are dynamic lags in the weapon drive, in the antenna drive, and in other mechanical and electrical elements within the system. The study of these effects, discussed in system configuration, involves consideration of the various ways in which the components of a weapon-control system can be arranged.

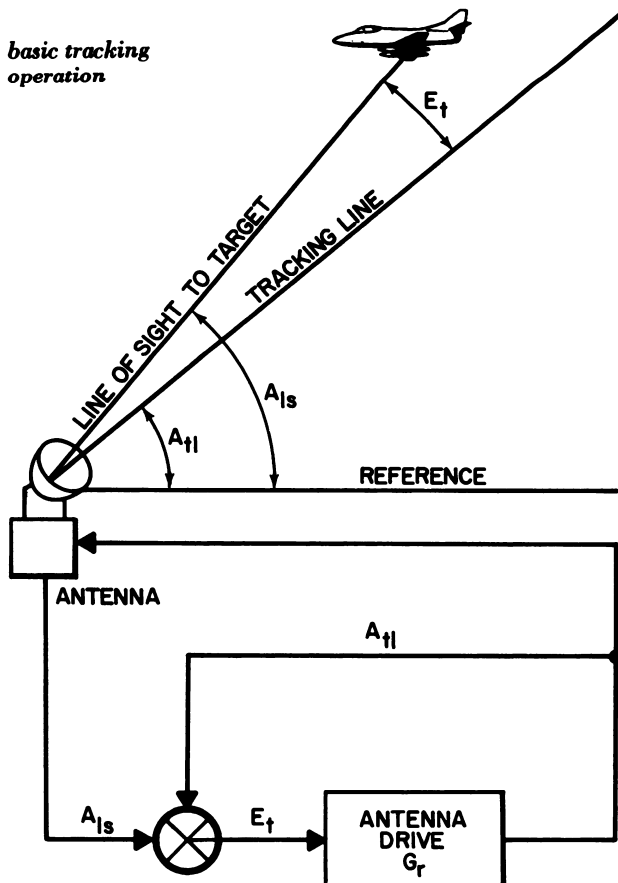
SENSOR NOISE

Tracking and prediction responses are fundamentally limited by sensor noise. To maintain a reasonable level of inherent sensor noise transmission to the weapon line, it is necessary to limit the lag time of the tracking response, the lag time of the prediction response, and the settling time of the tracking response to a ramp input. On the other hand, it is desirable that these responses be as fast as possible to reduce errors in target tracking and to limit the duration of the lock-on transients. Therefore, a fundamental compromise must be made in system design.

TRACKING

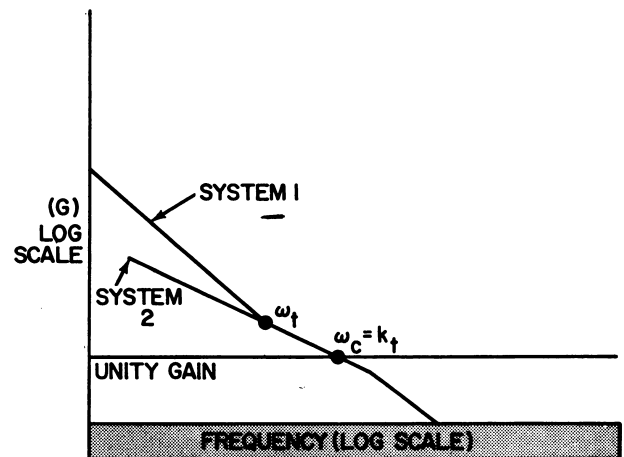
Angles considered in the tracking operation are the line-of-sight angle, A_{ls} , the tracking-line angle, A_{tl} , and the tracking error, E_t . The basic tracking operation and related angles are shown in the illustration. The tracking-error transducer, shown as a radar, produces the tracking error. This tracking error is applied to a control device (an antenna drive) which moves the tracking line in such a manner that the tracking error is minimized. The actual control device between E_t and A_{tl} may sometimes be simply an antenna drive, but in some complex systems it may be a very complicated multi-loop control system. Regardless of the type of control system, a transfer function between E_t and A_{tl} can always be defined. This transfer function, represented as G_t , defines the tracking operation. Let us now examine G_t in detail to determine its frequency-response requirements and how these requirements are related to the transient response.

basic tracking operation



One of the most important dynamic characteristics of the transfer function of a feedback-control loop is its bandwidth. The gain-crossover frequency, defined earlier in this chapter, determines the bandwidth. Since the asymptote-crossover frequency is simpler to calculate, it will be used in the remainder of this discussion instead of the actual gain-crossover frequency.

The major characteristics of the loop-transfer function G_t for a tracking loop are shown in the illustrated frequency response plot for System 2. This transfer function has an asymptote-crossover frequency, ω_c , which is designated as K_t and has a double integration of low frequencies. The upper break frequency in this double integration is defined as ω_t . A double integration is required in a tracking loop to achieve a high velocity constant. It is usually not necessary that this double integration hold at zero frequency (that is, the transfer function can break back to a single integration at very low frequencies), but there must be double integration for a significant frequency region below gain crossover to realize a high velocity constant. To demonstrate the effect of this double integration within the tracking loop, a plot of a similar transfer function is shown with a single integration at low frequencies and this plot is designated as System 1. A comparison of the transient responses for Systems 1 and 2 shows the effect of the low-frequency double integration in the tracking loop.



The unit step function response is illustrated for Systems 1 and 2. This illustration shows that the step responses for both systems rise to within 63 percent of the final value in the same time. The time for a transient response to rise to 63 percent is designated the rise time. It can be shown that the rise time for the step response of a feedback-control loop is approximately equal to the reciprocal of its gain-crossover frequency. Thus, the rise time for the tracking step response is roughly equal to $1/K_t$, which is the reciprocal of the asymptote-crossover frequency.

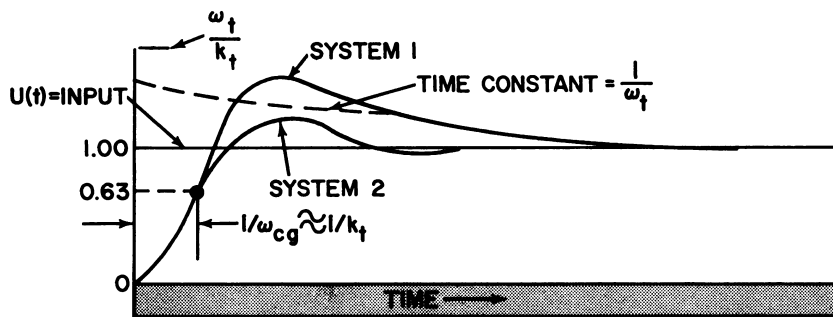
To obtain the fastest tracking response, it is necessary that the rise time of the step response of the tracking loop be as short as possible. This in turn requires that the gain-crossover frequency be large. On the other hand, the gain-crossover frequency of the tracking loop is basically limited by sensor noise. To maintain a reasonable amount of transmission of sensor noise to the tracking line, it is necessary that the bandwidth (that is, the gain-crossover frequency) of the tracking loop be maintained rather low. Generally, in airborne weapon-control systems, the gain-crossover frequency of the tracking loop must be limited to 1 cycle or less because of sensor noise. This rigid limitation on the gain-crossover frequency places a very rigid limitation on the rise time of the step response. For a gain-crossover frequency of 1 cycle (2π radians per second), the rise time of the step response is about $1/2\pi$ second.

The ramp response is another very important transient response used in describing the system performance in following target inputs. Most target courses approximate constant-velocity inputs to the system, and the tracking loop generally experiences a ramp input at the

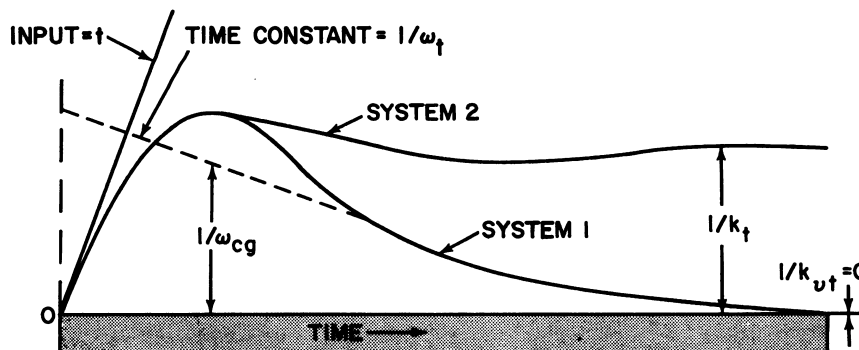
time the antenna first locks onto a moving target. If the target changes its angular velocity abruptly, the effect is a ramp input to the system.

The unit ramp response for Systems 1 and 2 is illustrated. Both systems have essentially the same maximum error for a unit ramp input, and this maximum error is approximately equal to the reciprocal of the gain-crossover frequency. Since the gain-crossover frequency is basically limited by sensor noise, there is a strict lower limit to the maximum error to a ramp. This maximum error can be reduced only by increasing the bandwidth. It is possible to decrease the steady-state error to a very small value by adding a low-frequency integral network. For example, the double integration in System 1 reduces the error for a ramp from $1/K_t$ to a final value of zero. The tail in the error response for System 1 essentially follows an exponential, as shown by the dotted curve. The time constant of this exponential is closely equal to $1/\omega_t$, where ω_t represents the upper break frequency of the integral network. Thus, the frequency ω_t essentially defines the setting time of the ramp response.

To achieve a fast lock-on transient in the system, the ramp error response must settle to zero as quickly as possible. A fast settling of the ramp response requires a high value for the frequency ω_t . On the other hand, the limitation that sensor noise places on the gain-crossover frequency, K_t , limits the value for ω_t that can be achieved with adequate stability. The double integration of System 1 adds phase lag in the region of the gain-crossover frequency K_t , and the higher its upper break frequency ω_t is with respect to K_t , the greater is the phase lag it adds at gain crossover, and hence the lower is the stability of the loop.



unit step function



unit ramp response

The unstabilizing effect of a double integration can be seen in another way by examining the step response. As illustrated previously, the step response of System 1 has more overshoot than the step response of System 2. This is because the double integration adds an exponential term (shown by the dotted curve) to the step response. The magnitude of this exponential is essentially proportional to the ratio ω_t/ω_c , and the higher ω_t is with respect to ω_c , the greater is the overshoot of the step response. It is also important to notice that the exponential adds a tail to the step response, and the time constant of this tail is approximately $1/\omega_t$.

Thus the tracking loop has two important frequency parameters that define its major transient characteristics. These are (1) the gain-crossover frequency, ω_c , and (2) the upper break frequency of the integration, ω_t . The gain-crossover frequency, ω_c , is fixed because of sensor noise to about 1 cycle, and ω_t in turn is fixed because it generally must be placed at least a factor of two or more below the frequency ω_c to achieve adequate stability. The gain-crossover frequency determines the rise time of the step response and the maximum error for the ramp response. The upper break frequency of the integration, ω_t , determines the settling time for the ramp response. The design of the tracking loop fundamentally represents a compromise between speed of response to target inputs and the transmission of sensor noise. To achieve faster tracking response, one must allow more sensor noise transmission, unless sensor noise can be reduced through radar improvement. This compromise is fundamental to all systems and is identical for all systems, regardless of the type of configuration the system may employ.

prediction

IDEAL PREDICTION

To consider the dynamic effect of prediction, assume that a target is following a circular course about the weapon station. Since ballistic and jump corrections have negligible effect upon system dynamics, the only computation necessary is the kinematic lead angle. The target course illustrated has an angular velocity designated as A_{ls} . To hit the target with a missile, the weapon line must lead the line of sight by the kinematic prediction angle, designated P' , that is

$$A_{wl} = A_{ls} + P'$$

During the time of flight, t_f , of the missile from the weapon station to the target, the target will have moved through the angle $t_f A_{ls}$. Therefore, the prediction angle, P' , is

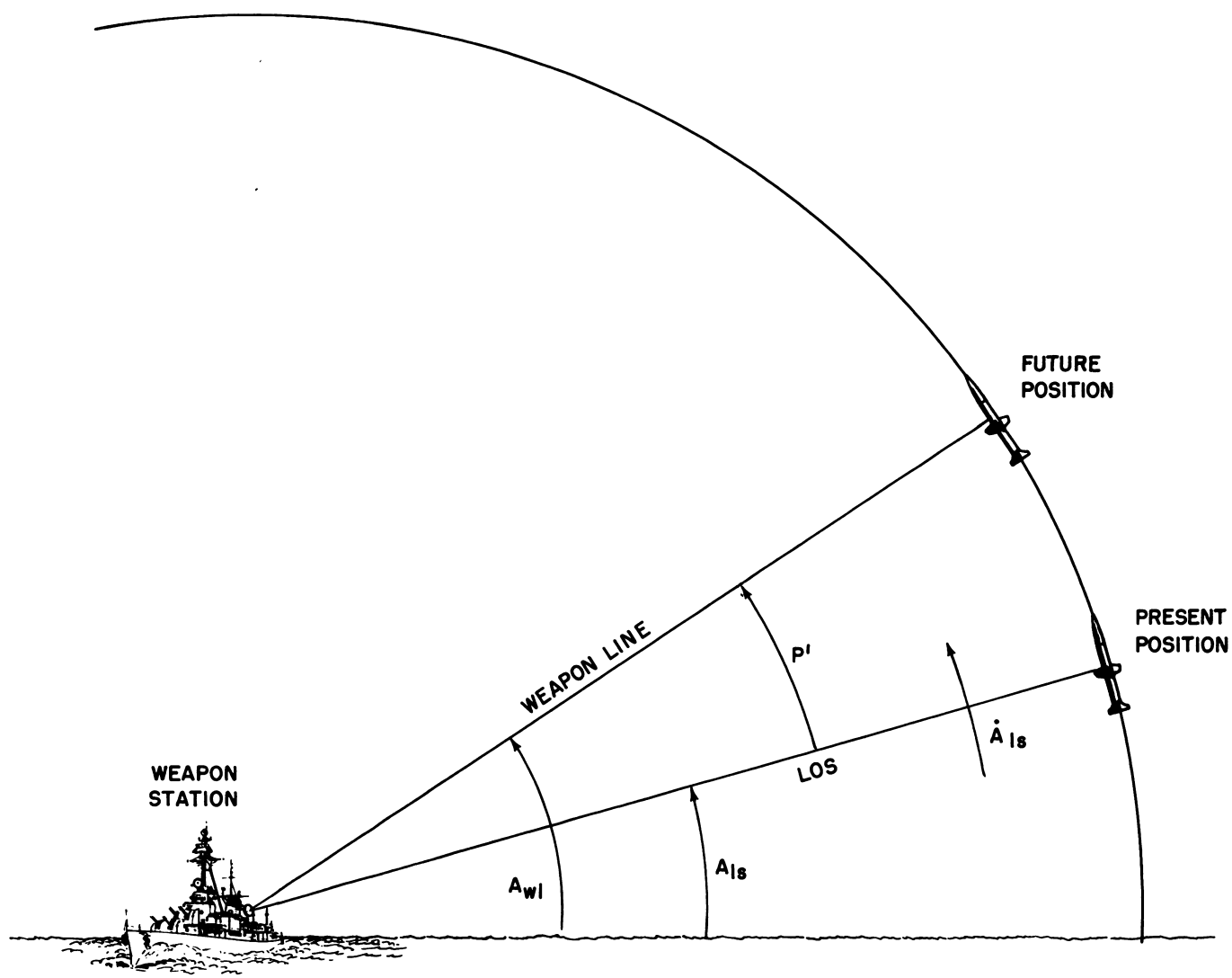
$$P' = t_f A_{ls}$$

This equation cannot be instrumented, since the line-of-sight angle is external to the system. However, the tracking-line angle, which is a measure of the line of sight, is developed in the system during the tracking operation. The actual equations instrumented in terms of the tracking-line angle are:

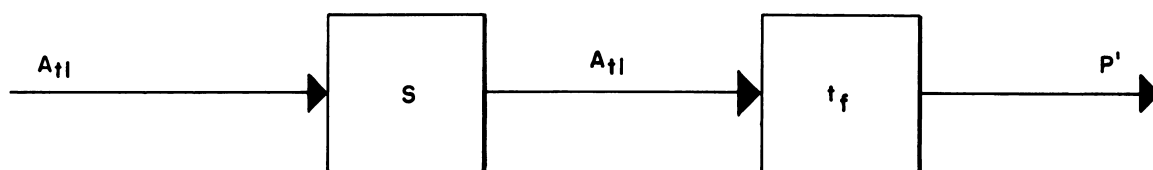
$$A_{wl} = A_{tl} + P'$$

$$P' = t_f A_{tl}$$

The weapon-line angle is equal to the tracking-line angle plus the prediction angle. The prediction angle is equal to the time of flight times the tracking-line velocity. The equation relating the tracking-line angle to the prediction angle may be obtained as shown in the illustration. The tracking-line angle is differentiated to produce the tracking-line velocity, A_{tl} . This differentiation is represented by the Laplace transfer operator, s , which is also often represented by the letter p .



circular course prediction



PREDICTION FILTERING

The prediction computation must always include some filtering. Since the tracking line angle contains sensor noise, which is accentuated in the differentiation process, a prediction filter is necessary. This prediction filter is designated by the transfer function G_f . The complete dynamic effect of the prediction computer can now be approximated by the equation:

$$P' = t_f G_f \dot{A}_{t1}$$

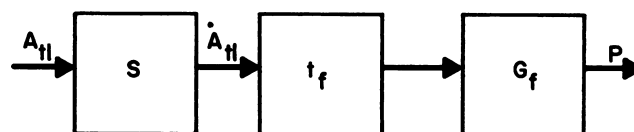
The prediction filter, G_f , limits the transmission of noise to the weapon line, but, at the same time, also limits the speed of response of the prediction action. The transfer function of G_f is a compromise between noise transmission and speed of response, just as was the setting of the tracking-loop parameters. This prediction filtering compromise is illustrated in detail in (a).

To determine the speed of prediction response, assume that there is a ramp of the tracking angle A_{t1} , which represents a step of tracking-line velocity \dot{A}_{t1} . Since the factor t_f varies slowly, it can be considered a constant, and the response of prediction angle to a step of tracking-line velocity is the same as the step response of the prediction filter G_f , shown in (b) of the illustration. This step response represents the response of prediction angle to a step of tracking-line velocity. Let us arbitrarily define the time for the step response to reach 63 percent of its final value as equal to $1/\omega_b$, where ω_b is arbitrarily called the bandwidth.

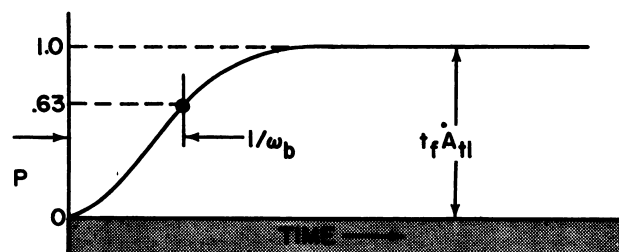
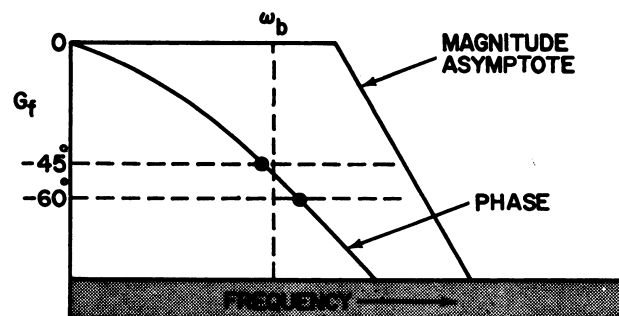
The frequency response for a typical filter function, G_f , is shown in (c) of the illustration. The frequency ω_b , defined previously, generally lies between the frequencies at which 45-degree and 60-degree phase lag occur. Therefore a basic relationship exists between the characteristics of the phase in the prediction-filter transfer-function and the rise time of the prediction-filter step response.

To provide adequate filtering of sensor noise, the prediction-filter transfer function must begin attenuating noise at a low frequency. This requirement results in a limit on the phase curve, thus restricting the value for ω_b that can be achieved with a given amount of sensor noise transmission. The limitation on ω_b in turn limits the rise time of the prediction response.

Therefore, to reduce the rise time of the prediction response to a step of tracking-line velocity, it is necessary that the bandwidth ω_b be increased, and consequently that the noise transmission be increased.



A) PREDICTION FILTERING

B) P FOR STEP OF \dot{A}_{t1} 

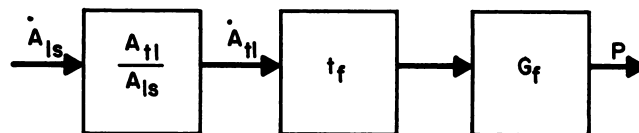
C) FILTER FREQUENCY RESPONSE

The previous illustration does not completely define the prediction operation, because the actual line-of-sight angle should be considered as the input signal rather than the tracking-line angle. The transient response of the complete prediction operation should be defined as the prediction response for a ramp of line-of-sight angle, rather than for a ramp of tracking-line angle.

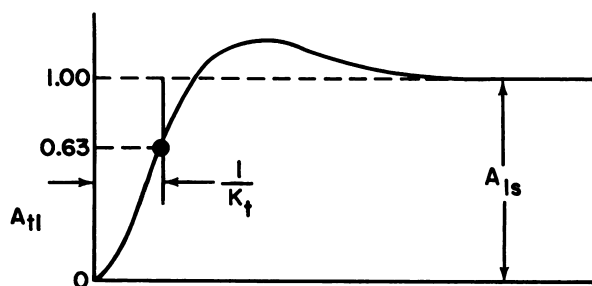
The total transmission between line-of-sight velocity \dot{A}_{ls} and prediction angle P is shown in (a) of the illustration. The illustration contains two low-pass filters in cascade: the tracking operation, with the transfer function A_{tl}/A_{ls} , and the prediction filter, with the transfer function G_f .

When two low-pass filters are cascaded, it can be shown that the rise time of the resultant step response is approximately equal to the sum of the rise times of the individual step responses of the two filters. Thus, the rise time for the overall prediction operation is essentially equal to the sum of the tracking-response rise time and the prediction-filter rise time. The step response of the tracking operation A_{tl}/A_{ls} , which was considered previously, is approximated in (b) of the illustration and its rise time is essentially equal to $1/K_t$, where K_t is the tracking-loop gain-crossover frequency. The step response for the complete prediction operation is approximately as shown in (c) of the illustration with a rise time equal to $1/K_t + 1/\omega_b$. Usually, the prediction-filter rise time is much longer than the tracking-response rise time, so that in practice the prediction filter primarily determines the speed of response of the prediction operation.

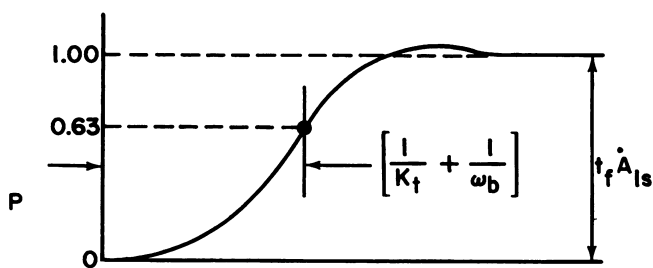
The relationship between the bandwidth of the prediction-filter function G_f and the speed of prediction response is fundamental for all weapon-control systems regardless of the type of instrumentation employed. Sometimes the prediction filter is instrumented directly and its transfer function is determined primarily by the characteristics of an electrical network in the computer. In other methods of instrumentation, it may not be as obvious. Nevertheless, the transfer function G_f applies in all cases and represents a fundamental system parameter. The characteristics of G_f are a compromise between prediction response and noise transmission due to prediction.



A) COMPLETE PREDICTION TRANSFER FUNCTION



B) A_{tl} FOR STEP OF A_{ls}



C) P FOR STEP OF \dot{A}_{ls}

WEAPON-LINE COMPUTATION

Weapon-line computation represents a method of prediction instrumentation for which the transfer function for the prediction filter G_f is not obvious. In weapon-line computation, the lead angle is computed from the weapon-line velocity rather than from the tracking-line velocity. An important reason for computing in this manner is that in some cases it is difficult, if not impossible, to measure the tracking-line velocity directly and an indirect measure of tracking-line velocity is obtained from the weapon line.

The prediction-filter function for a weapon-line computer can be found by computing the resultant transfer function between tracking-line angle and prediction angle. To show this, consider the equations solved in weapon-line computation. They are based upon the fundamental lead-angle computer equations that have been presented:

$$A_{wl} = A_{tl} + P$$

$$P = t_f \dot{A}_{tl}$$

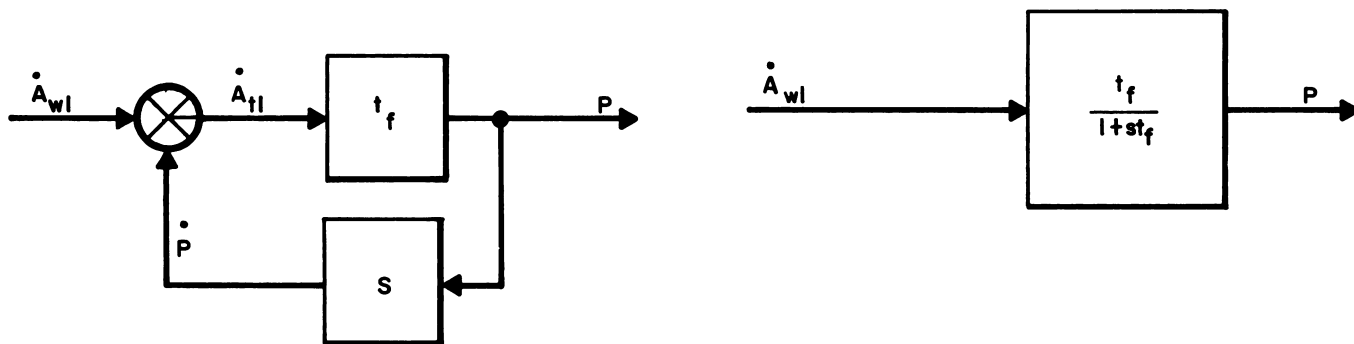
Differentiating the first equation gives

$$\dot{A}_{tl} = \dot{A}_{wl} - \dot{P}$$

Thus, the tracking-line velocity can be found indirectly by measuring the weapon-line velocity and subtracting from it the rate-of-prediction \dot{P} . This tracking-line velocity can be used in the second equation, to compute the prediction angle, P .

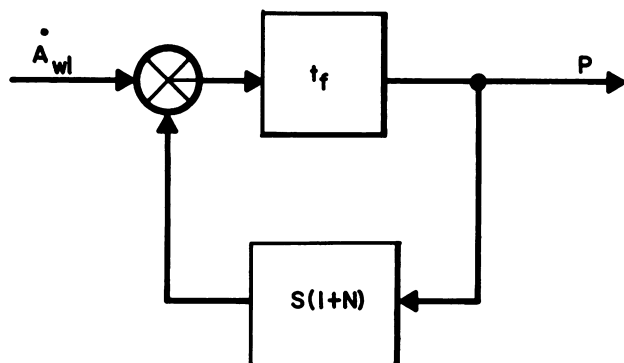
The basic weapon-line computation procedure is illustrated by a block diagram. The weapon-line velocity represents the input in the lead-angle computation and the prediction angle is the output. The prediction angle is differentiated to obtain the rate-of-prediction \dot{P} . This is subtracted from the weapon-line velocity to obtain the tracking-line velocity, which in turn is multiplied by the time of flight to obtain the prediction angle. Thus,

the weapon-line computer essentially employs a feedback loop. The resultant transfer function between weapon-line velocity and prediction angle can be condensed in the form of a single block as illustrated.



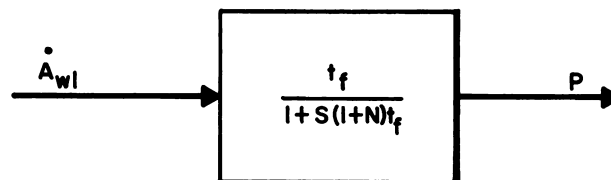
basic weapon line computation

The diagram for basic weapon-line computation is incomplete in that prediction filtering is not included in the computation. A simple way to add prediction filtering to weapon-line computation is to increase the amount of the prediction-rate feedback, so that the resultant block diagram is as illustrated. The rate feedback is increased by the ratio $1 + N$, where N is called the stability number and is frequently set at about 0.2.



weapon line computation with prediction filtering

The feedback loop of this block diagram can again be expressed as a single cascade block as illustrated.

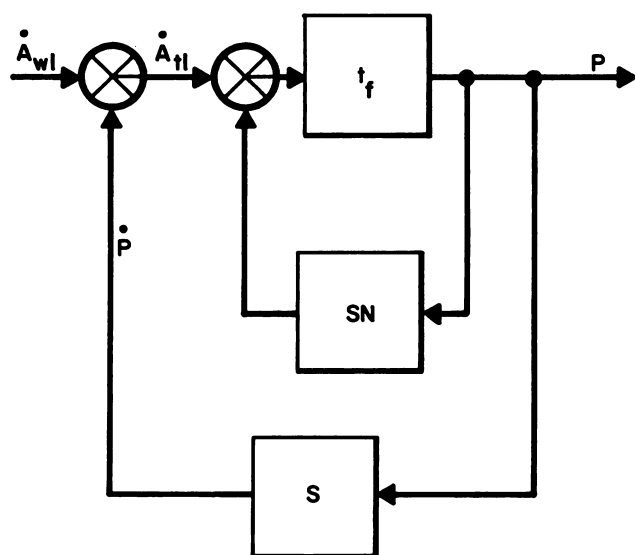


This diagram is finally condensed, as illustrated, showing that the increased prediction-rate feedback has effectively added a prediction filter, G_f , with the transfer function

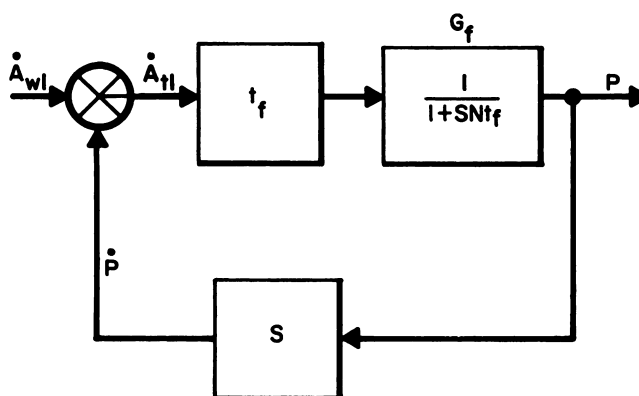
$$G_f = \frac{1}{1 + sNt_f}$$

This filter is a single-order lag with a time constant proportional to time of flight.

To determine the equivalent prediction filtering, the block diagram is further modified to the form illustrated in which the prediction-rate feedback is shown as two parallel paths.



parallel prediction-rate feedback



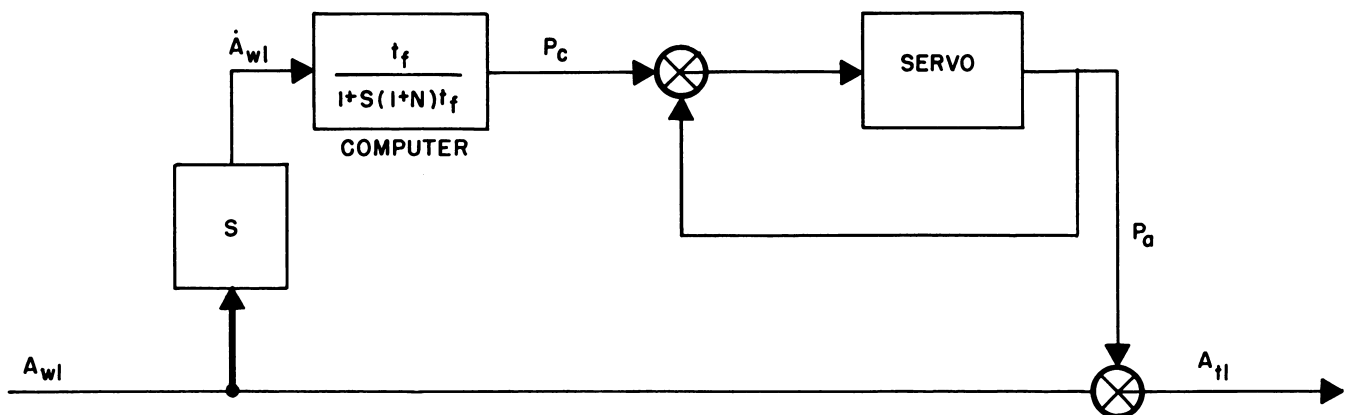
Very often there is a servomechanism such as an antenna loop following the computer to develop the actual prediction angle, as illustrated in the block diagram. In the diagram, the computer output is called the computed prediction angle, P_c , and the output from the prediction servomechanism is called the actual prediction angle, P_a , which is the true angle between the weapon line and tracking line; that is,

$$P_a = A_{wl} - A_{tl}$$

The transfer function for the computer itself is assumed to be the same as was shown previously:

$$\frac{P_c}{\dot{A}_{wl}} = \frac{t_f}{1 + sNt_f}$$

The total transfer function between weapon-line velocity and actual prediction angle, then, is equal to the last equation multiplied by the transfer function P_a/P_c for the prediction servo.

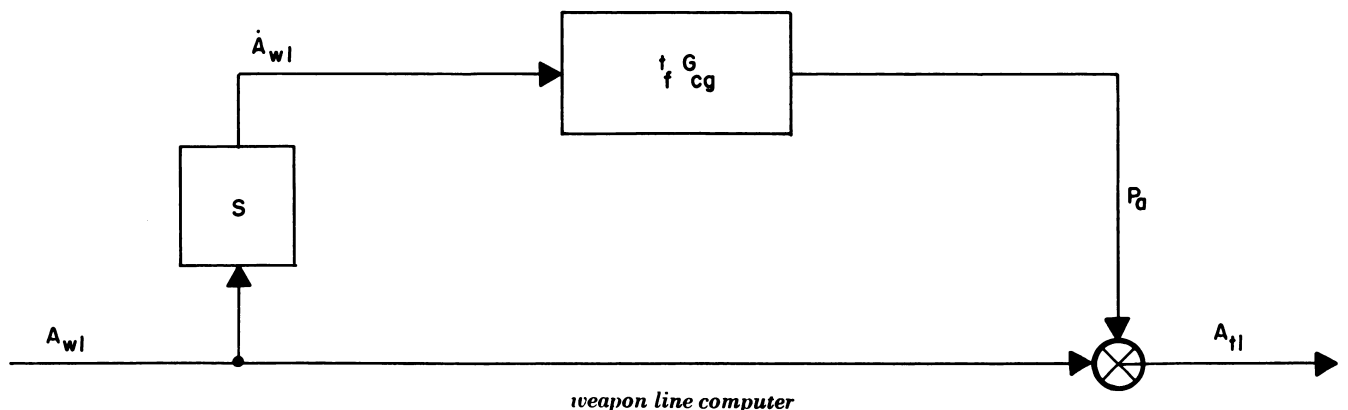


To compute the prediction-filter function, G_f , for the weapon-line computer illustrated, the block diagram is condensed as shown. The complete weapon-line computer transfer function is designated G_{cg} , which is defined so that

$$P_a = (t_f G_{cg}) \dot{A}_{wl}$$

Hence G_{cg} is equal to

$$G_{cg} = \frac{1}{t_f} \frac{P_a}{\dot{A}_{wl}} = \frac{P_a/P_c}{1 + s(1+N)t_f}$$



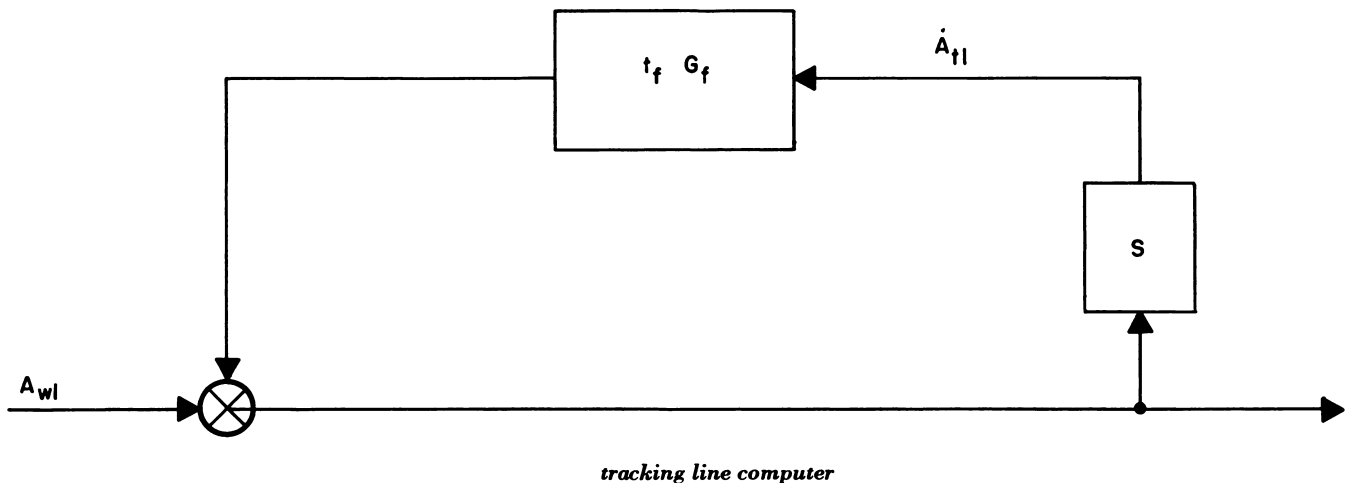
The equivalent block diagram for a tracking-line computer is also illustrated. By equating the transfer functions between A_{gl} and A_{tl} for the weapon line and tracking line computers, it can readily be shown that the equivalent prediction-filter function for the weapon line computer is:

$$G_f = \frac{G_{cg}}{1 - s t_f G_{cg}}$$

Thus, the prediction-filter function G_f still has significance when weapon-line computation is employed. The same design compromise still must exist between noise transmission and prediction response. By computing the equivalent prediction filtering by means of the previous equation for G_f , the essential elements of this compromise can be expressed in a convenient manner for optimum design. It is much easier to evaluate the dynamic effect of prediction filtering for a weapon-line computer by examining the equivalent prediction-filter function G_f than by considering the weapon-line computer transfer-function G_{cg} directly.

Frequently, when weapon-line computation is employed, the importance of the prediction servomechanism dynamics is not adequately appreciated. It is often felt that if the bandwidth of the prediction servomechanism is significantly greater than the bandwidth of the resultant prediction filter G_f , the dynamics of the servomechanism can be neglected, and the resultant prediction filter can be assumed to be merely the value given below.

$$G_f = \frac{1}{1 + s N t_f}$$



However, by means of the equation for G_f discussed previously, it can be shown readily that the dynamics of the prediction servomechanism generally causes G_f to be underdamped, and that a tremendously wide bandwidth is often needed in the prediction servomechanism to overdamp G_f . This tendency toward an underdamped prediction-filter function is important in closed loop applications, since it affects stability. When weapon-line computation is employed, it is essential that the dynamics of the prediction servomechanism be considered carefully.

WEAPON STATION MOTION

A weapon-control system that is mounted on a moving weapon station must perform the important job of stabilizing the tracking line on the target, against the disturbing action of the weapon station motion. Although this operation is sometimes performed by the tracking loop in its normal action of keeping the tracking line pointed on the target, weapon-control systems often have an additional gyro loop that performs most of the space-stabilization operation.

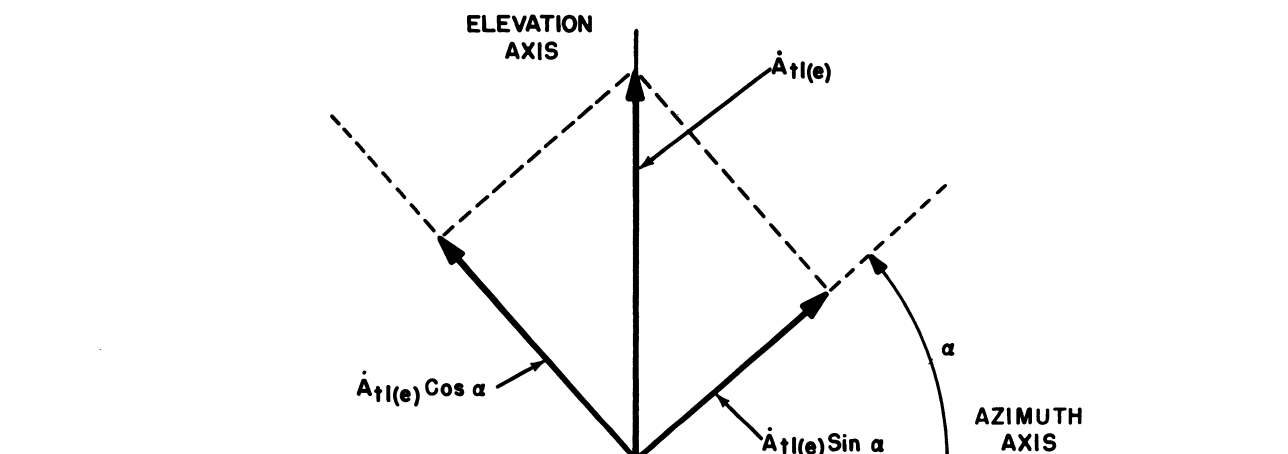
The stabilization operation of a weapon-control system generally operates only in two axes (azimuth and elevation) so that it maintains the tracking line pointed on the target, but does not stop the tracking-member axes from rotating about the tracking line. This rotation of the tracking (or antenna) axes about the tracking line is called cross roll about the tracking line. Similarly, a rotation of a gun carriage or a missile about the weapon line is called cross roll about the weapon line. Cross roll about the tracking line is primarily caused by motion of the weapon station, although a small amount of cross roll can result from the normal motions of the antenna azimuth gimbal as the antenna follows the target.

Cross roll about the tracking line can produce errors because of the dynamic lags in the parts of the system involved in tracking and prediction. The solid lines in the illustration represent the initial orientation of the elevation and azimuth axes of the antenna. Assume that the target is moving at a constant angular velocity in the elevation plane and that the antenna, following the target, is rotating about its elevation gimbal with an elevation tracking-line velocity components $\dot{A}_{tl(e)}$. Then assume that the tracking axes (antenna gimbals) suddenly rotate about the tracking line by the angle α , so that they are oriented as shown by the dashed lines. For the antenna to maintain the same velocity with respect to inertial space, it must develop a velocity component about its azimuth gimbal equal to $\dot{A}_{tl(e)} \sin \alpha$. It must also reduce the angular velocity about its elevation gimbal to the value $\dot{A}_{tl(e)} \cos \alpha$. Therefore, the two separate axes of the weapon-control system experience a changing target-rate input. Since the tracking and prediction operations are performed separately in the

two axes, this effective change in target velocity produces dynamic errors, unless a means of correction is provided.

The easiest way of offsetting the dynamic errors caused by cross roll is to use a three-axis tracking device, roll stabilized about the tracking line. This maintains the orientation of the tracking axes fixed with respect to inertial space, so that cross roll about the tracking line cannot occur. On the other hand, the requirement for a three-axis gimbal system to carry the stabilization gyros can greatly increase system complexity.

Another method of handling cross-roll errors is to add correction signals for compensation. By measuring the cross-roll velocity with a third gyro, the cross-roll errors can be computed, and equal and opposite corrections added to the system. This method may be much simpler than the actual roll stabilization of the tracking member. Since the cross-roll errors are generally relatively small, it may be completely adequate, even with relatively crude computational techniques.

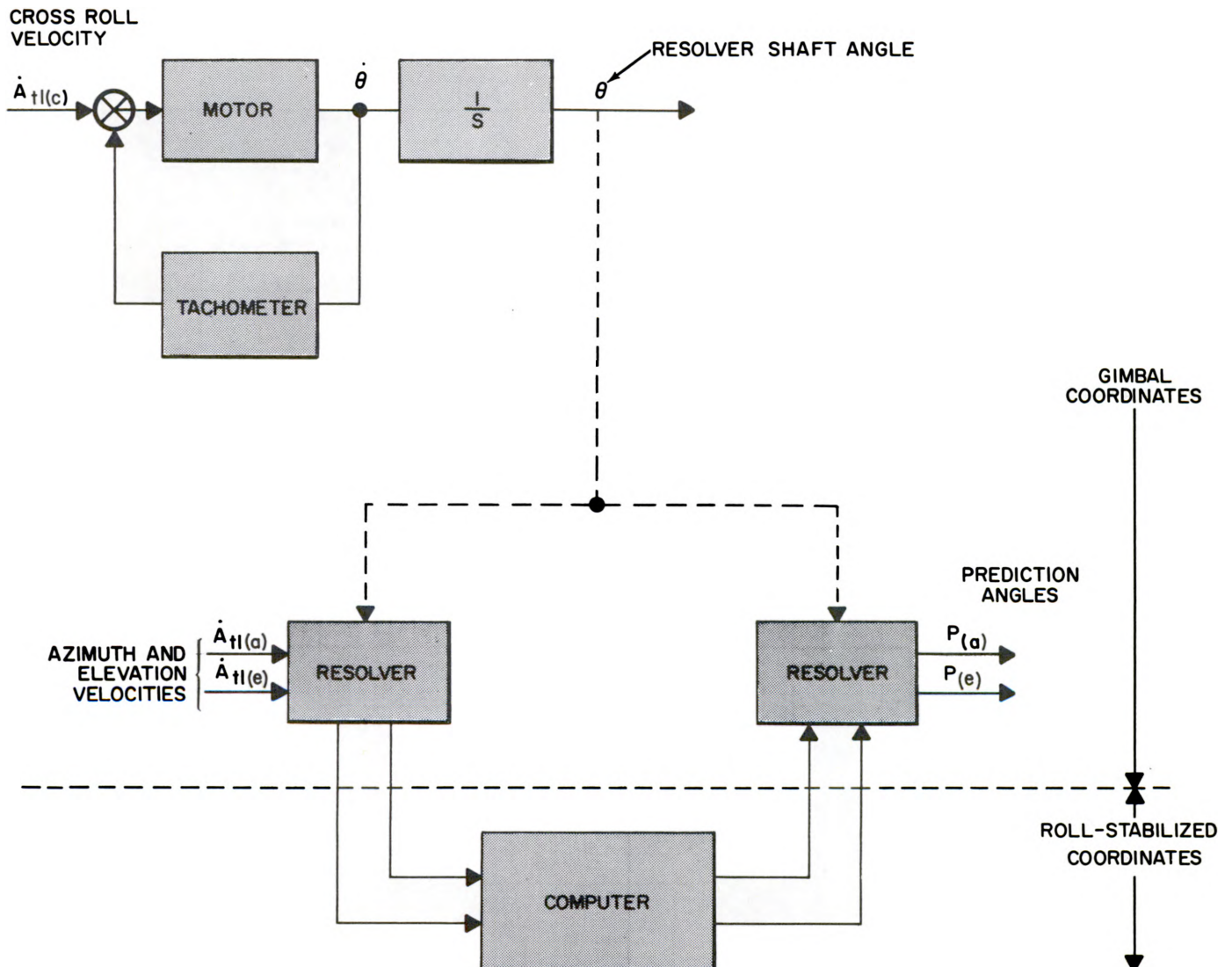


prediction cross roll

A simple means of roll stabilizing the prediction operation to counteract the cross-roll errors caused by the prediction lag is illustrated. A small rate gyro is placed on the antenna. This gyro delivers a signal proportional to the cross-roll angular velocity, $\dot{A}_{tl}(c)$, about the tracking line. This cross-roll velocity signal is fed to an instrument servo mechanism, which contains a motor-driven tachometer and a pair of resolvers. The tachometer signal is fed back around the motor to close a rate-feed-back loop. The gain in the tachometer feedback path is adjusted to match the angular velocity from the rate gyro in such a manner that there is a one-to-one correspondence between the angular cross-roll rotation about the tracking line and an angular rotation of the resolvers on the instrument-servo mechanism shaft. When the azimuth and elevation tracking-line velocity signals $\dot{A}_{tl}(a)$ and $\dot{A}_{tl}(e)$ are applied into one of the resolvers, two velocity components result that are roll-stabilized with respect to inertial space. These are fed to the computer. Then, all the prediction fil-

tering can be performed upon these roll-stabilized components without cross-roll error. The prediction angle components from the computer are applied to the second resolver and resolved back into azimuth and elevation coordinates.

When the illustrated technique is employed, it is not necessary that the cross-roll gyro have very high degree of accuracy nor is it necessary that the tachometer feedback gain be adjusted critically. A relatively large amount of drift can be tolerated in the rate gyro and the instrument servo mechanism, because such drift has the effect only of a rather slow cross roll about the tracking line. Even with rather fast cross-roll rotations, cross-roll errors are generally quite small. In fact, cross-roll errors resulting from the lags in prediction and tracking were considered sufficiently small to be completely neglected in some short range weapons systems. However, in long range systems and beam rider guided missile systems cross roll errors cannot be ignored.



tracking cross roll

Although the previously illustrated technique effectively roll-stabilizes the prediction operation, it does not affect the cross-roll error attributable to the lag in tracking. A simple method for effectively roll stabilizing the tracking operation is to employ the following technique. As discussed previously, when the tracking-line axes are rotated by the angle α , the following angular velocity component has to be developed about the azimuth axis to maintain the same velocity with respect to inertial space:

$$\Delta \dot{A}_{tl(a)} = -\dot{A}_{tl(c)} \sin \alpha$$

If the angle α is very small, $\sin \alpha$ may be approximated by α ; and the required change in azimuth velocity is:

$$\Delta \dot{A}_{tl(a)} = -\dot{A}_{tl(e)} \alpha$$

Now, assume that the axes are rotating about the tracking line with a constant cross-roll angular velocity $\dot{A}_{tl(c)}$. Then the incremental rotation α over a small increment of time, Δt , is

$$= \dot{A}_{tl(c)} \Delta t$$

Combining the last two equations gives for the required change in azimuth angular velocity over the time increment Δt :

$$\Delta \dot{A}_{tl(a)} = -\dot{A}_{tl(e)} \dot{A}_{tl(c)} \Delta t$$

Therefore, to maintain a fixed velocity with respect to inertial space, the total change required in azimuth angular velocity over an interval of time may be expressed by the integral:

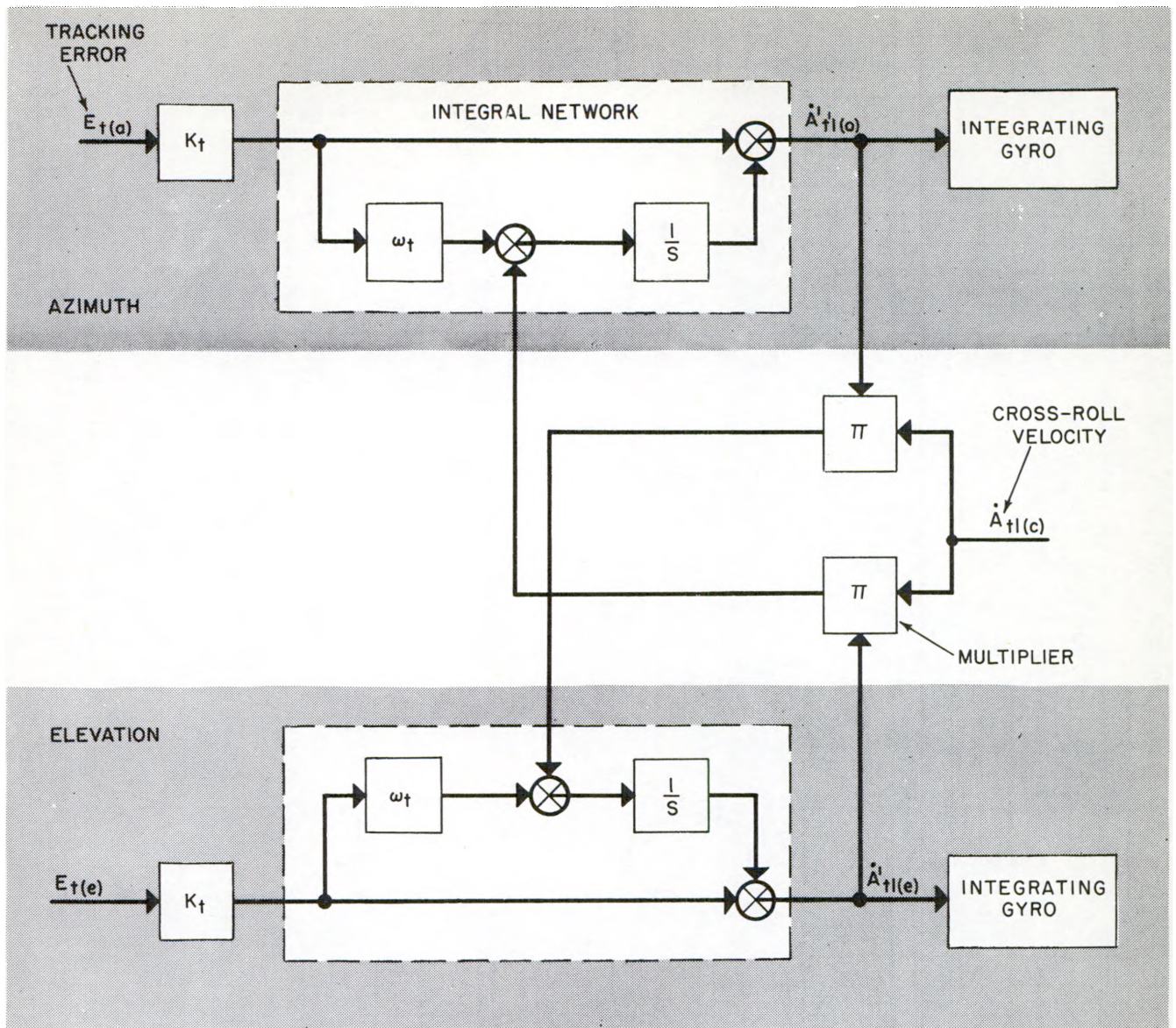
$$A_{tl(a)} = - \int dt \dot{A}_{tl(e)} \dot{A}_{tl(c)}$$

Multiplying the elevation and cross-roll tracking-line velocities and integrating the product, a correction signal is obtained which can be added to the azimuth tracking-line angular velocity to compensate for the effect of cross roll on the tracking operation. A similar equation can be computed for the correction that must be added to the elevation angular velocity.

The block design shows how a weapon-control system can be instrumented to use the last equations as a means of correcting for the cross-roll tracking error. This figure contains the tracking amplifier sections for the azimuth and elevation axes of a weapon-control system using integrating gyros. The signals going into the gyros, $\dot{A}'_{tl(a)}$ and $\dot{A}'_{tl(e)}$, which represent measures of the azimuth and elevation tracking-line rates, $\dot{A}_{tl(a)}$ and $\dot{A}_{tl(e)}$, are used both for lead computation and for cross-roll correction. There is a rate gyro on the tracking line that furnishes a cross-roll angular velocity signal $\dot{A}_{tl(c)}$. This signal is multiplied by the elevation velocity signal $\dot{A}'_{tl(e)}$, as indicated by a block labeled π , and the product is subtracted at the input to the integrator in the azimuth channel. The gain in the path is adjusted so that the resultant change in the azimuth tracking-line rate-command signal $\dot{A}'_{tl(a)}$ is equal to the value given by the last equation. In like manner, the azimuth velocity signal $\dot{A}'_{tl(a)}$ is multiplied by the cross-roll rate, and the product is added at the input to the integrator in the elevation channel.

The illustrated technique can theoretically provide an exact correction for cross-roll tracking errors. However, because of inaccuracy in the cross-roll gyro and inexact matching of the various paths, some error will occur. Since the overall correction is small, it should generally be relatively simple to maintain the resultant error at a negligible point.

Although a pure integration is employed which could be supplied by a mechanical integrator, the same technique could be employed with an undercompensated electrical integral network.



weapon-line cross roll

When weapon-line computation is employed, cross-roll can cause a third type of cross-roll error which, when it occurs, is generally much worse than the cross-roll errors attributable to the lags in tracking and prediction. As was shown previously, the tracking-line rate is computed in weapon-line computation by subtracting the rate of prediction from the weapon-line rate. This computation of tracking-line rate is exact, however, only if the weapon-line rate and prediction rate lie in the same plane. For reasonably small prediction angles, the errors are generally quite small, except when cross roll about the weapon line occurs. The effect of a cross-roll angular velocity $\dot{A}_{gl}(c)$ about the weapon line is to change the computed value of elevation tracking-line rate by the amount

$$\Delta \dot{A}_{tl}(e) = \dot{A}_{wl}(c) \sin P_d$$

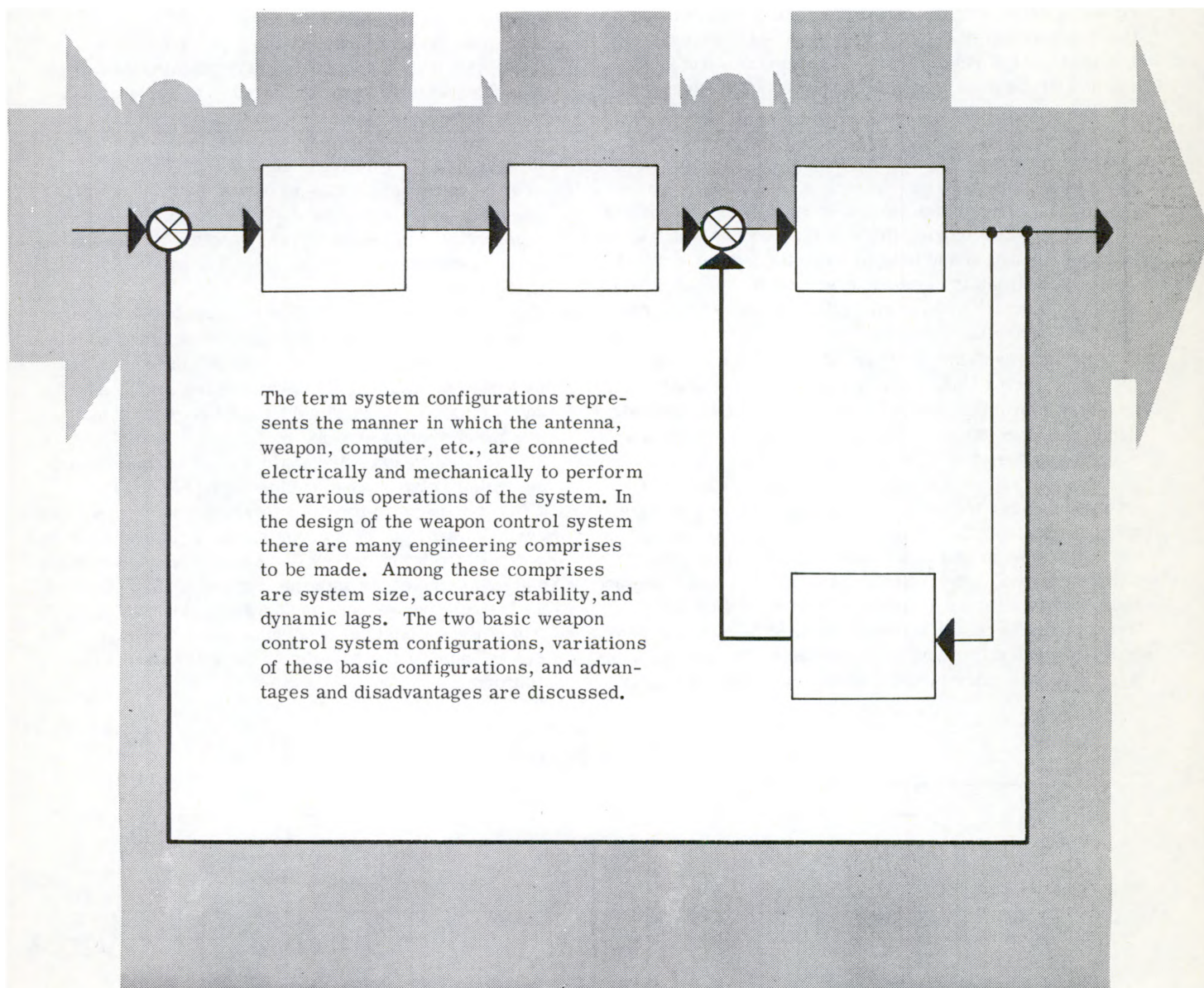
and to change the computed azimuth tracking-line rate by

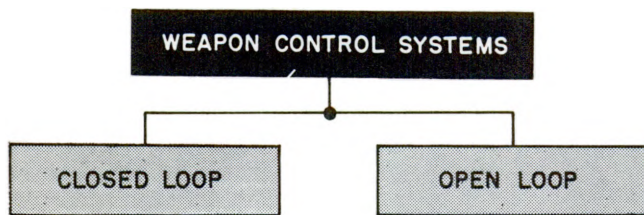
$$\Delta \dot{A}_{tl}(a) = \dot{A}_{wl}(c) \cos P_d \sin P_e$$

Where P_d and P_e are the deflection and elevation prediction angles, respectively. (The equations assume that the elevation antenna carriage is carried within the azimuth carriage.) These corrections represent the components of the weapon-line cross-roll velocity vector resolved along the azimuth and elevation axes of the tracking line.

To compensate for the cross-roll errors attributable to weapon-line computation, the negatives of the two rate components in these equations are effectively inserted into the computer to correct the computed components of tracking-line velocity.

SYSTEM CONFIGURATIONS





The basic breakdown of weapon control system configurations is illustrated. The two main categories are: the open-loop and the closed-loop configurations. In open-loop systems, the antenna and weapon gimbal frames are mounted separately on a common weapon platform. In closed-loop systems, the antenna gimbal frame is mounted on the weapon carriage and the relative angle between the antenna and weapon line is measured directly by a transducer on the antenna gimbal. The closed-loop category has a static-accuracy advantage over the open-loop category in that there is one rigid mechanical gimbal system between antenna and weapon line and thus there is little chance for bore-sight error resulting from mechanical flexure existing between the two. In addition, the data links in a closed-loop system make use of the relative angles between antenna and weapon line, rather than the total weapon-line and tracking-line angles. The handling of relative angles tends to increase static accuracy and simplify coordinate-conversion computation. On the other hand, the closed-loop category has dynamic disadvantage. Since the antenna rides on the weapon carriage, the operations of prediction and moving the weapon tend to disturb the tracking operation and tend to make the system unstable. Sometimes dynamic performance must be degraded below what could be achieved in an equivalent open-loop system in order to maintain stability.

In open-loop systems, motions of the weapon cannot affect the tracking line, and consequently the operations of tracking, prediction, and moving the weapon can be independent of each other. Thus this category is dynamically superior to the closed-loop category. However, from a static-accuracy viewpoint, it is inferior, because there is much greater tendency for boresight error resulting from mechanical flexure between the separate units. The data links must handle total weapon line angles and antenna angles, rather than relative angles, and therefore they require greater accuracy of instrumentation. Therefore, the choice between the closed-loop and open-loop categories involves a compromise. In open-loop systems, it is more difficult to achieve static accuracy,

but simpler to achieve a rapid response than in closed-loop systems; the reverse situation prevails in open-loop systems.

open loop system

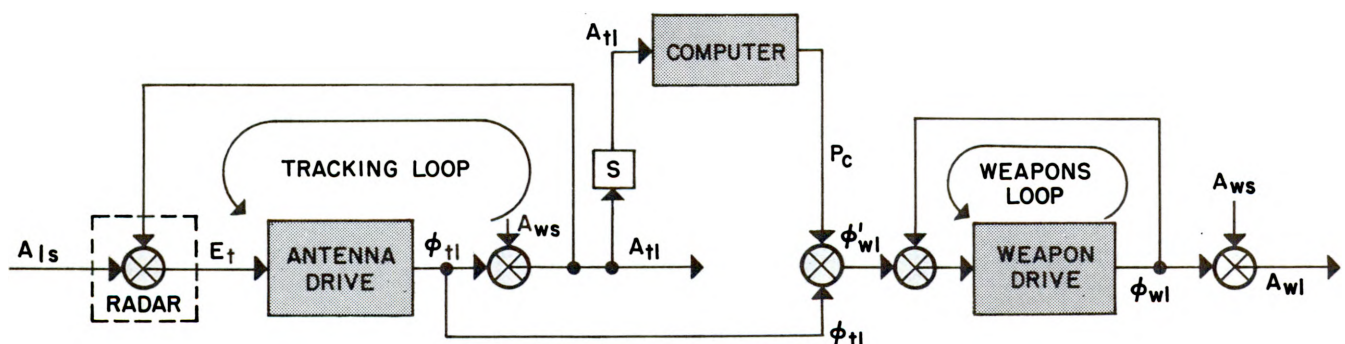
A block diagram of a basic open-loop system is provided. The system performs three fundamental operations: tracking, prediction, and positioning the weapon. These operations are performed by the tracking loop, the computer, and the weapon loop respectively, as shown in the illustration.

It is important when considering the various angles of the system, to differentiate between a relative angle measured between two members and an angle measured with respect to inertial space. The symbol A represents an angle measured with respect to inertial space, and the symbol ϕ represents an angle measured with respect to the weapon station. Consider the weapon station as the platform on which the antenna and weapon gimbals are mounted, and the angle of the weapon station with respect to inertial space as A_{ws} .

The radar measures the tracking error E_t between the line-of-sight angle A_{ls} and the tracking-line angle A_{tl} . The tracking error E_t is fed to the antenna drive, which develops a relative displacement ϕ_{tl} of the antenna with respect to its gimbal. As shown in the illustration, the relative angle ϕ_{tl} must be added to the weapon-station angle A_{ws} to obtain the total inertial-space tracking-line angle A_{tl} . The angle A_{tl} is differentiated (often by a gyro) to obtain the inertial-space tracking-line angular velocity, which is applied to the computer.

The computer computes the prediction angle P_c (a relative angle). The prediction angle is summed with the relative tracking-line angle ϕ_{tl} to obtain the desired weapon-line angle ϕ'_{wl} , also a relative angle. This addition is often performed by a differential synchro in a synchro-data system.

The desired weapon-line angle ϕ'_{wl} is compared with the actual relative weapon-line angle ϕ_{wl} and the difference between the two represents the weapon-loop error signal E_w . This error signal actuates the weapon control loop which brings the actual weapon-line angle ϕ_{wl} into coincidence with the loop input ϕ'_{wl} . Control of this relative weapon-line angle ϕ_{wl} results in a control of the inertial space weapon-line angle A_{wl} , since A_{wl} is equal to ϕ_{wl} plus the angle of the fixed weapon-platform A_{ws} .

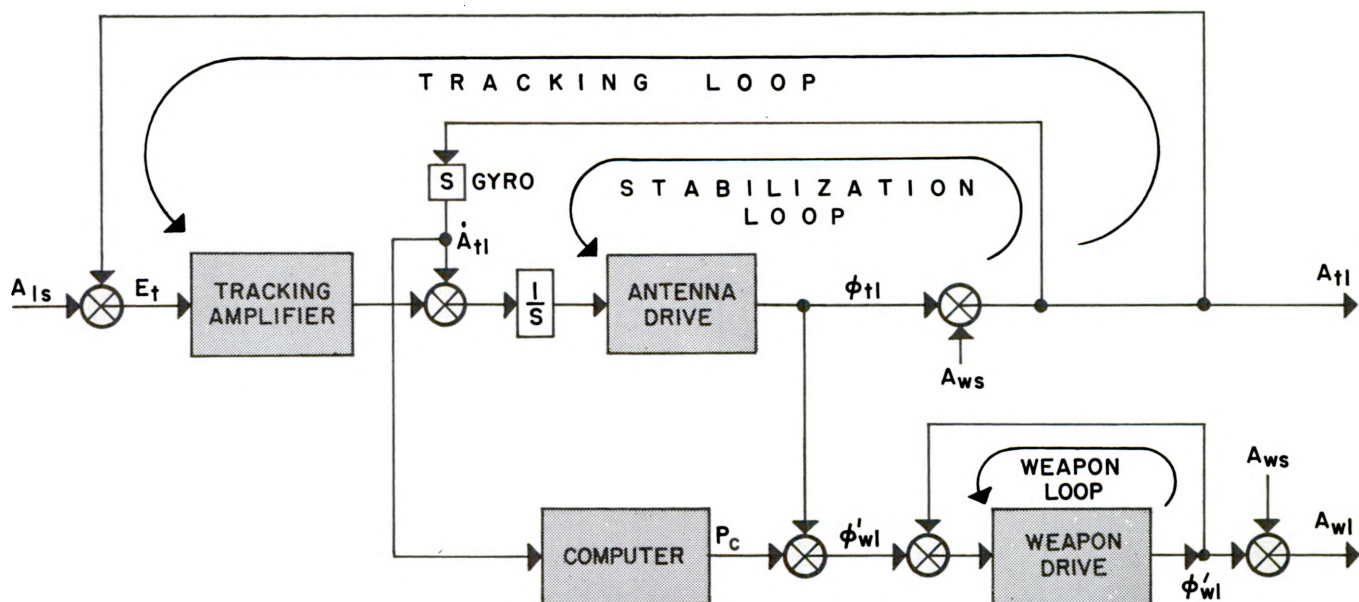


The distinction between relative- and inertial-space angles is quite important when one considers the effect of weapon-station motion. For land-based applications, the weapon-station is truly fixed, but for airborne and shipboard applications, the weapon station moves. This motion of the weapon station must be considered as a disturbing input to the system.

The illustration shows that the weapon-station angle A_{ws} represents a disturbance which is an input to the tracking loop; hence, any weapon-station motion can produce tracking error unless the motion is at a much lower frequency than the gain-crossover frequency of the tracking loop. However, for many airborne appli-

cations this is not so, especially when the aircraft performs a maneuver. As has been shown, the gain-cross-over frequency of the tracking loop must be quite low (about 1 cycle) in order to limit radar-noise transmission. Although such a bandwidth is adequate for tracking target motions, it is often not adequate for offsetting the effect of weapon-station motion.

Consequently, an additional control loop is required to perform the independent operation of stabilizing the tracking line with respect to the disturbance caused by weapon-station motion. This loop is formed by placing a gyro on the antenna and feeding its signal around the antenna drive.



The open loop system, modified by the addition of a gyro stabilization loop to the antenna drive is illustrated. The gyro signal \dot{A}_{t1} that is applied to the computer provides the stabilization signal, and this signal is applied into the input of the antenna drive amplifier. It can be seen that the computer and weapon-loop section shown in the previous illustration are identical with those shown here.

The block diagram of the tracking section is shown with heavy lines. Two loops are closed about the antenna: the stabilization loop and the tracking loop. The stabilization loop is closed around the weapon-station input A_{ws} , and the gain-crossover frequency of this loop is made as high as the dynamics of the antenna drive and gyro permit. The tracking loop is closed outside the stabilization loop and adjusted for low bandwidth to provide proper filtering of the tracking signal. In practice, a higher gain-crossover frequency is designed for a gyro-stabilization loop than for a tracking loop. When a wide separation of bandwidths exists, the two loops operate essentially independently of each other. The gyro stabilization loop response is faster than the tracking loop response and the error signal in the gyro stabilization loop resulting from an input from

the tracking amplifier is quite small. The stabilization-loop return-signal \dot{A}_{t1} is approximately equal to the tracking-amplifier output signal. Therefore, the stabilization loop has the effect in the tracking loop of a pure integration between the tracking amplifier output and the tracking-line angle A_{t1} , because the tracking-amplifier output must be closely equal to the tracking-line velocity \dot{A}_{t1} . Consequently, by adjusting the tracking-amplifier gain, the tracking-loop bandwidth can be adjusted independently of the stabilization-loop bandwidth. Since there is a differentiation in the gyro transfer function and only one integration in the antenna drive, a further integration must be added to obtain a net integration within the stabilization loop. This additional integration is represented by the factor $1/s$ in the stabilization loop. Without this additional integration, the stabilization loop would not have adequate gain. It is not necessary that the additional integration operate at frequencies below the gain crossover of the tracking loop because the tracking loop controls the antenna position at low frequencies. Consequently, the additional integration can be approximated adequately by an electrical lag network, provided it has a break frequency below the tracking-loop gain-crossover frequency.

antenna-drive tracking

If the antenna is moved from the weapon station and mounted directly on the weapon carriage, a closed-loop antenna-drive tracking configuration is produced. The main effect on the previous block diagram of moving the antenna gimbal from the weapon station to the weapon carriage is to replace the signal A_{ws} fed into the stabilization loop by the signal A_{wl} , thus producing a feedback from the weapon loop down into the stabilization loop.

The resultant block diagram with the antenna mounted in closed-loop fashion is shown. The feedback of the weapon line A_{wl} into the stabilization loop is shown by a heavy line. With the antenna mounted in a closed loop, a new symbol must be employed for the relative tracking-line angle, since the antenna gimbal now moves with respect to the weapon line rather than with respect to the weapon station. Accordingly, the relative angle of the antenna with respect to the weapon is designated θ_{tl} , so that the inertial-space tracking-line angle A_{tl} is equal to θ_{tl} plus the weapon-line angle A_{wl} :

$$A_{tl} = \theta_{tl} + A_{wl}$$

On the other hand, in the analysis of prediction action, the actual prediction angle P_a was defined as

$$P_a = A_{wl} - A_{tl}$$

Comparing these equations shows that the relative tracking-line angle θ_{tl} is equal to the negative of the actual prediction angle P_a . Hence, when the relative angle θ_{tl} is added to the computed prediction angle P_c , the result is equal to the error in prediction E_p , as illustrated in the diagram. That is,

$$E_p = P_c + \theta_{tl} = P_c - P_a$$

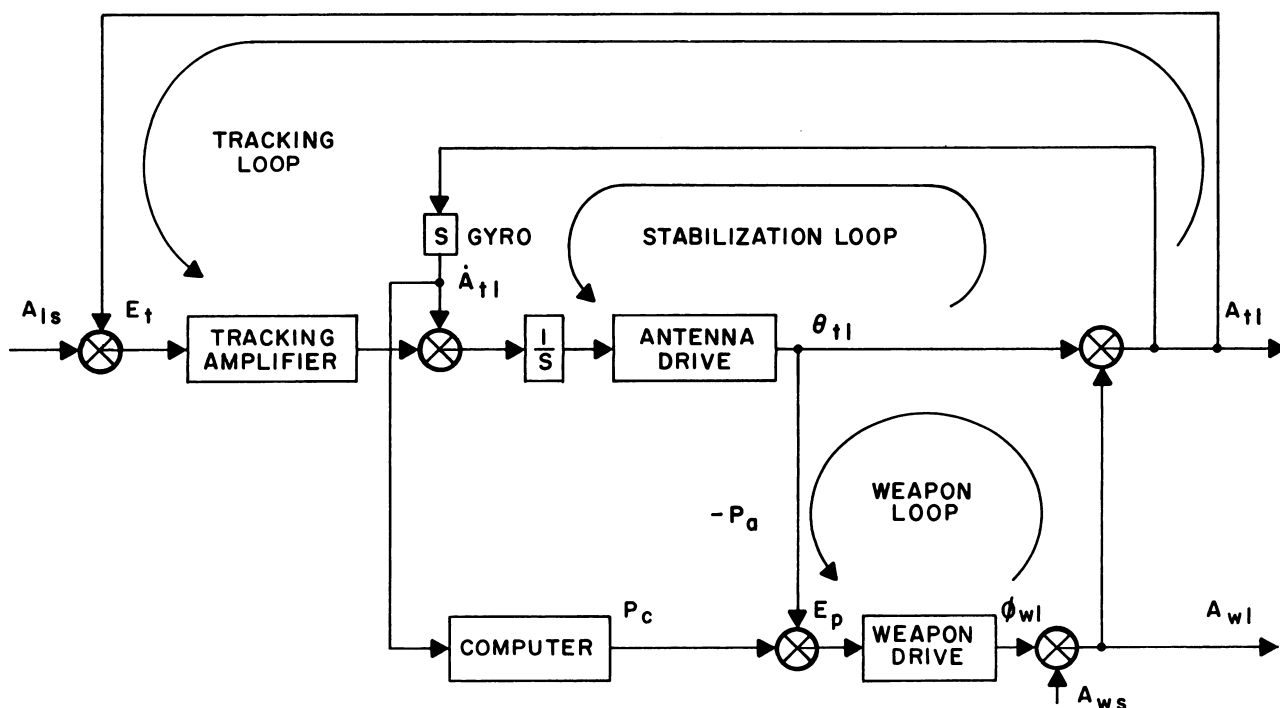
The error in prediction, E_p , is fed to the weapon line drive, which positions the weapon line to reduce E_p to zero.

It should be noted that there is no direct position loop closed around the weapon drive in the illustrated

antenna-drive tracking system as there was in the open-loop system previously discussed. The weapon loop is now closed about the antenna gimbal so that weapon-line position is measured with respect to the antenna rather than with respect to the weapon station. If there is a sufficiently fast stabilization-loop closed around the antenna, the antenna appears to the weapon to be fixed in inertial space and therefore it makes little difference whether the weapon angle is measured with respect to the stabilized antenna or with respect to a fixed platform. A comparison of the two systems shows that the block diagram of the antenna-drive tracking system is the same as that for the open-loop system except for changes in the weapon loop. The main stability requirement of the antenna-drive tracking system can be deduced by simple reasoning. As illustrated, the weapon-line angle A_{wl} represents a disturbing input to the stabilization loop, just as weapon-station motion represented a disturbance for the open-loop system. Besides producing tracking error, however, this disturbance in the closed loop case can lead to instability. For adequate stability, the gain-crossover frequency of the stabilization loop must be sufficiently greater than that of the weapon loop in that disturbances caused by the motion of the weapon line have a negligible effect on the tracking line. Therefore, for stable operation, the weapon loop and stabilization loop must operate essentially independently of one another.

To illustrate this stability requirement quantitatively, in an experimental antenna-drive tracking turret, a gain-crossover frequency in the antenna stabilization loop of about 15 cycles is sufficient to allow a 3.5-cycle gain-crossover in the weapon loop.

The prediction signal also tends to disturb the tracking line; hence, another requirement for stability is that the bandwidth of the prediction signal be much less than that of the stabilization loop. However, for flexible turrets of the present-day bomber, it appears that if



the stabilization-loop bandwidth is wide enough to support an adequate weapon-loop bandwidth, it should be perfectly adequate to support as high a prediction bandwidth as sensor noise permits.

It is always desirable that the weapon loop be as fast as the weapon-line drive dynamics permit to minimize errors in offsetting weapon-station motion and in tracking the target. No matter how fast or accurate the other parts of the system are, a target hit cannot be scored unless the weapon platform can be kept pointed in the target direction. On the other hand, with antenna-drive tracking, the bandwidth of the weapon loop must be much less than that of the stabilization loop, which in turn is limited by the dynamics of the antenna drive and the gyro. Thus, the antenna-drive tracking configuration requires that the antenna drive and gyro have a much wider bandwidth than the weapon drive. For present-day airborne turrets, the requirement on the gyro is readily achieved, and the antenna drive requirement is generally not excessive, since the antenna can be much smaller than the weapon carriage and therefore can be much stiffer.

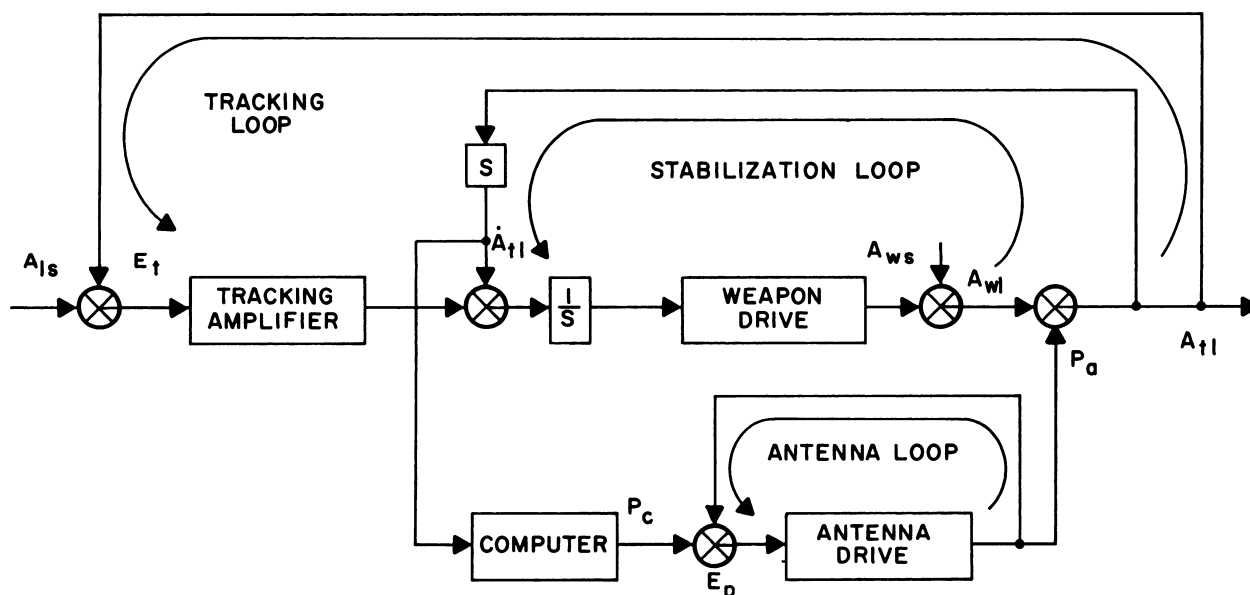
weapon-drive tracking

In the antenna-drive tracking configuration, the antenna drive performs the operations of tracking and stabilizing the tracking line against disturbing motions of the platform on which the antenna is mounted. On the other hand, with the antenna mounted in a closed loop it is possible to perform these operations by moving the platform on which the antenna is mounted - - i.e., the weapon carriage - - rather than by moving the antenna drive. This method is employed in the weapon-drive tracking configuration.

To understand the operation of the weapon-drive tracking configuration, assume first that the prediction angle is zero. The antenna and weapon then move as a rigid member and the whole system tracks like a large antenna. The tracking error is fed to a stabilization loop closed

about the weapon-line drive, and controls the weapon line to point at the target, just as it controlled the antenna to point at the target in the open-loop configuration. To develop a prediction angle, a servo loop closed around the antenna drive slowly moves the antenna in its gimbals. This results in a tracking error which displaces the weapon line in the reverse direction by the same angle, bringing the tracking line back on the target. Thus, the weapon tracks as it did before, but the weapon line is biased off from the tracking line by the prediction angle.

A block diagram for the weapon-drive tracking system is shown. The heavy lines show the tracking- and space-stabilization operations which are performed by the weapon drive. Note that the elements in this section of the block diagram are the same as in the heavy-line section of the antenna-drive tracking block diagram previously shown, except that the antenna drive is replaced by the weapon drive. Although the weapon drive is performing the task of space stabilization, the stabilization gyros are still placed on the antenna, because it is the tracking line that is actually being stabilized. The gyros used for stabilization also supply the tracking-line velocity signal for the computer. The computer develops the computed prediction angle P_c and feeds this to a servo loop closed around the antenna gimbal. This servo loop develops the actual prediction angle P_a by displacing the tracking line from the weapon line by the angle $-P_a$, and slaves P_a to follow P_c . The stability requirements again can be deduced quite simply. As the illustration shows, the prediction angle P_a represents a disturbing input to the stabilization loop. It is being fed into the stabilization loop in parallel with the weapon-station angle A_{ws} , which also represents a disturbing input, but unlike A_{ws} , the prediction signal P_a also can produce instability. For adequate stability, the stabilization loop must be much faster than the prediction signal - - that is, the gain-crossover frequency of the stabilization loop must be much greater than the bandwidth of the prediction filter.



In contrast to the antenna-drive tracking configuration, the stabilization-loop bandwidth for the weapon-drive tracking configuration is limited by the dynamics of the weapon drive rather than by the dynamics of the antenna drive. Since the weapon carriage must be much larger than the antenna, the achievable stabilization-loop gain-crossover frequency in the weapon-drive tracking configuration generally must be significantly less than that of the antenna-drive tracking configuration (provided adequate care is taken in the design of the antenna drive). This limitation on the stabilization-loop bandwidth places a strong limitation on the allowable bandwidth for the prediction filter. On the other hand, system interaction places no limitation on the weapon-loop bandwidth when weapon-drive tracking is employed. In fact, it is desirable that the bandwidth of the control loop around the weapon (i.e., the stabilization loop), be as wide as the weapon-drive dynamics permit to achieve better stability.

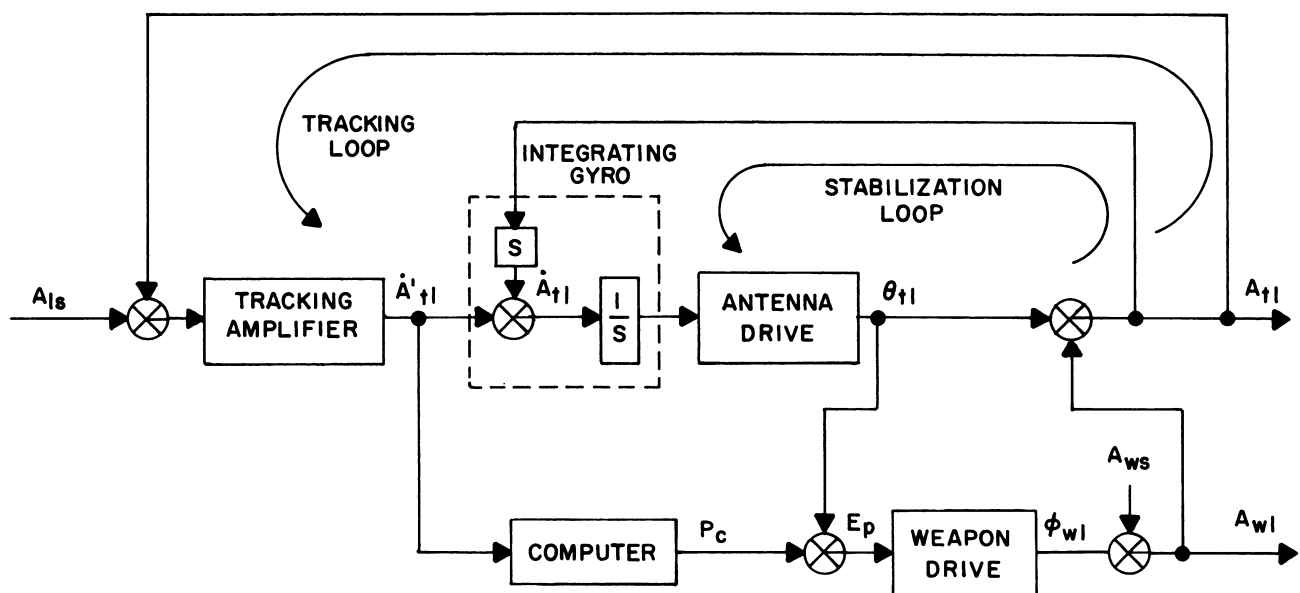
Integrating gyros

When integrating gyros instead of rate gyros are employed, the form of the block diagram of the system is changed, but this change in the block diagram does not materially affect the system operation. As an example, consider the block diagram which shows an antenna-drive tracking system with integrating gyros. The gyro action produces a torque on the gyro gimbal shaft proportional to tracking-line angular velocity. Since the gimbal shaft is restrained only by a viscous fluid, this torque drives the gyro gimbal shaft at a rate proportional to the tracking-line angular velocity. Thus, an angular displacement of the tracking line with respect to inertial space produces a proportional angular displacement of the gimbal shaft in its bearings. A microsyn coupled to the gyro gimbal measures its position and thereby gives a measure of angular displacement of the tracking line with respect to inertial space. There is also a torque motor that supplies a second torque input to the gyro-gimbal shaft. The torque motor signal and the

angular rate signal are summed as torques on the gimbal shaft, so that the microsyn signal is proportional to the integral of the sum of the torque motor input and the tracking-line angular rate. The resultant transfer function for the integrating gyro is illustrated by the dashed box.

The integrating gyro does not directly supply a signal proportional to the tracking-line rate for the computer input. However, when the gyro is operated in a wide bandwidth stabilization loop, the signal being fed into the gyro torque motor is an excellent measure of tracking-line rate. This signal is designated \dot{A}'_{t1} and may be considered to be the desired tracking-line velocity. For reasons of accuracy and, in closed-loop systems, for reasons of stability as well, the stabilization-loop bandwidth should be much greater than the prediction-filter bandwidth, and consequently the error in the stabilization loop should be negligible at frequencies passed by the computer. Thus, so far as the computer is concerned, the stabilization-loop input \dot{A}'_{t1} is equal to the stabilization-loop return signal, which is the tracking-line angular rate \dot{A}_{t1} . This condition holds statically as well as dynamically, since the integrating gyro provides a pure integration and has extremely small dead-space. The gyro torque motor input, therefore, is an excellent measure of tracking-line rate for the computer.

Use of the integrating gyro distinctly changes the form of the system block diagram, because the input to the computer is fed from a point external to the stabilization loop rather than from a point within the stabilization loop. This change in block diagram form does not in itself appreciably change the system dynamic performance. However, dynamic performance can usually be improved with an integrating gyro, because its bandwidth is generally wider than that of an equivalent rate gyro, thus allowing a higher gain-crossover frequency in the stabilization loop.



computing from weapon line

For weapon-drive tracking systems it is possible to compute the lead angle from the weapon-line rate instead of from the tracking-line rate without affecting the system dynamics, since the operation of tracking is performed by the weapon-line drive. On the other hand, when weapon-line computation is instrumented in practice, the equivalent prediction filtering is often somewhat underdamped, and consequently it can adversely affect dynamic operation. However, as was shown previously, it is possible to obtain in a weapon-line computer the same filtering as in a given tracking-line computer. If this is done, the dynamic operation should be the same for the two computers in a weapon-drive tracking configuration.

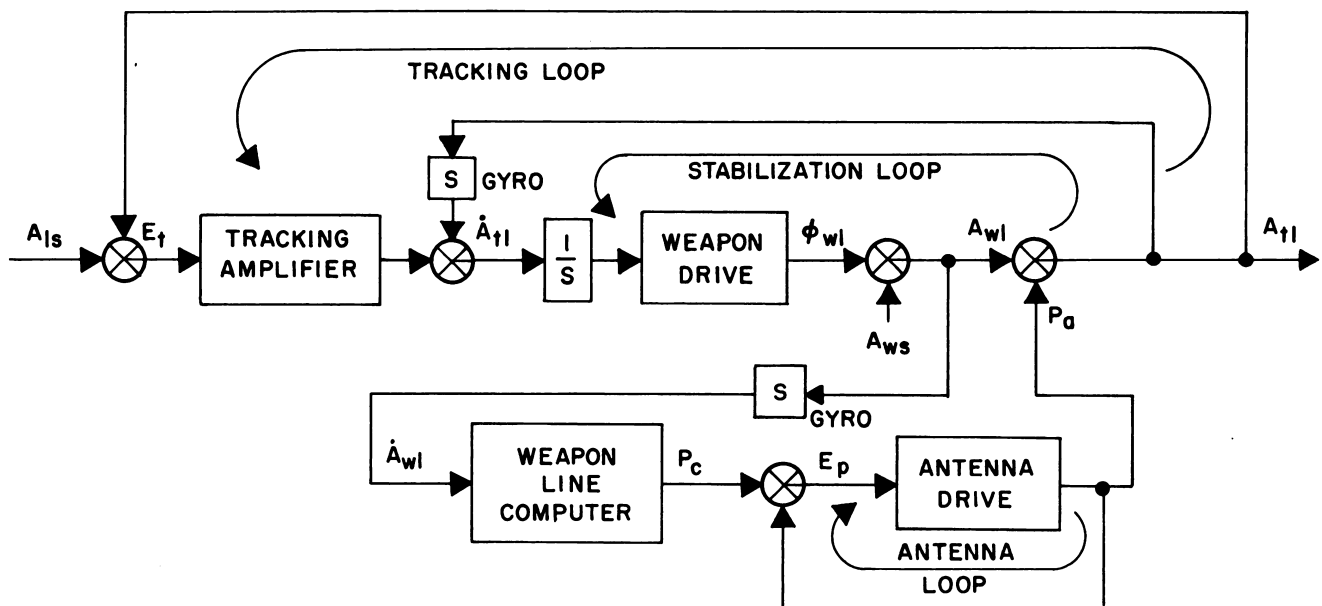
In an antenna-drive tracking system, the dynamic performance might be seriously degraded if the lead angle were computed from the actual weapon line, because stable operation would require that the prediction-filter bandwidth be much less than the gain-crossover frequency of the weapon loop. Since the weapon loop must be much slower than the stabilization loop, this requirement is more severe than the requirement that prediction bandwidth be less than the gain-crossover frequency of the stabilization loop, which applies when tracking-line computation is used. Similarly, for open-loop systems, computing from the weapon-line rate could severely limit prediction response if the weapon loop is slow.

It should be emphasized that there is an important distinction between using weapon-line computation and actually computing from the weapon-line rate. A number of open-loop and antenna-drive tracking systems use weapon-line computation. These systems compute from the rate of an auxiliary member in the system pointed in the direction of the desired weapon line, instead of computing from the rate of the actual weapon line, and the actual weapon line is slaved to follow this auxiliary member. For such cases, the

prediction bandwidth is limited not by the gain-cross-over frequency of the weapon loop but rather by the gain-crossover frequency of the stabilization loop around the auxiliary member, which generally can be sufficiently high to allow adequate prediction response. An early and very successful application of computation from the weapon line was in an optical weapon-drive tracking system, the Mark 14 gunsight, for shipboard antiaircraft. This gunsight was designed to be mounted on a manually operated gun. By means of rotating mirrors, the operator's sight (i.e., the tracking line) was displaced from the weapon line by the negative of the prediction angle; so that when the target was being tracked smoothly in the operator's sight, the gun was leading the target by the required prediction angle. Gyros in the sight case measured the weapon-line angular rate, and the computer used weapon-line computation to compute the prediction angle. The use of tracking-line computation for this application would have been very difficult, because gyros could not have been mounted on the delicate mirrors that defined the tracking line.

Another system used for fixed-gun fighters, and quite similar dynamically to the Mark 14 gunsight is illustrated. The sight is mounted on the aircraft, which is the weapon carriage, and the pilot performs tracking by steering the aircraft.

It should be noted that although these sights are gyro sights and are used in closed loops, neither has a gyro stabilization loop. Stabilization of the tracking line is performed entirely by the manually operated tracking loop. For manually operated weapon-drive tracking systems, it appears that using the optical tracking loop for stabilization can be reasonably adequate, but experience has shown that gyro stabilization is necessary for automatic radar tracking systems. Probably one reason for this is that the human operator is able to modify his transfer function to maintain stability in the system despite the changing dynamic effect of the computer.



If gyro stabilization is to be used in a weapon-drive tracking system with weapon-line computation, it is the tracking line that requires stabilization, but the gyro need not necessarily be mounted on the antenna. The signal stabilizing each axis of the tracking line is obtained from a rate gyro mounted on the antenna or it is obtained by summing together the signals from a gyro mounted on the weapon carriage and a tachometer on the antenna gimbal, as illustrated. In the former case, the gyros supplying the computer signals were an integral part of the computing mechanism, which was mounted on the weapon carriage; hence, these gyros could not provide the stabilization signals. The latter, being a more recent system, employs integrating gyros mounted on the weapon carriage for computation and the same gyros also provide the stabilization signals. The closed-loop system just described differs from the weapon-drive tracking system described previously only in that the weapon-line rate \dot{A}_{wl} is fed into the computer rather than the tracking-line rate \dot{A}_{tl} . The integrating gyro on the weapon carriage provides the weapon-line rate signal for the computer and the signal for a stabilization loop closed around the weapon drive. This stabilization loop corrects for weapon-station motion but does not offset the disturbing action of prediction. To stabilize the tracking line against prediction, there is an additional prediction-rate signal P_a fed back around the weapon line drive. This signal is obtained from a tachometer on the antenna gimbal. Thus, there are two parallel signals fed around the weapon drive, and their sum represents a total negative feedback equal to the weapon-line rate minus the prediction rate, which is equal to the tracking-line rate. The system stabilizes the tracking line by feeding a tracking-line rate signal around the weapon drive, just as is done in the closed-loop system without the integrating gyro but obtains this rate signal in a slightly different manner. There is a dynamic difference between the system with and the one without the integrating gyro that can be important in certain applications. In the system without the gyro, the stabilization loop is closed around the

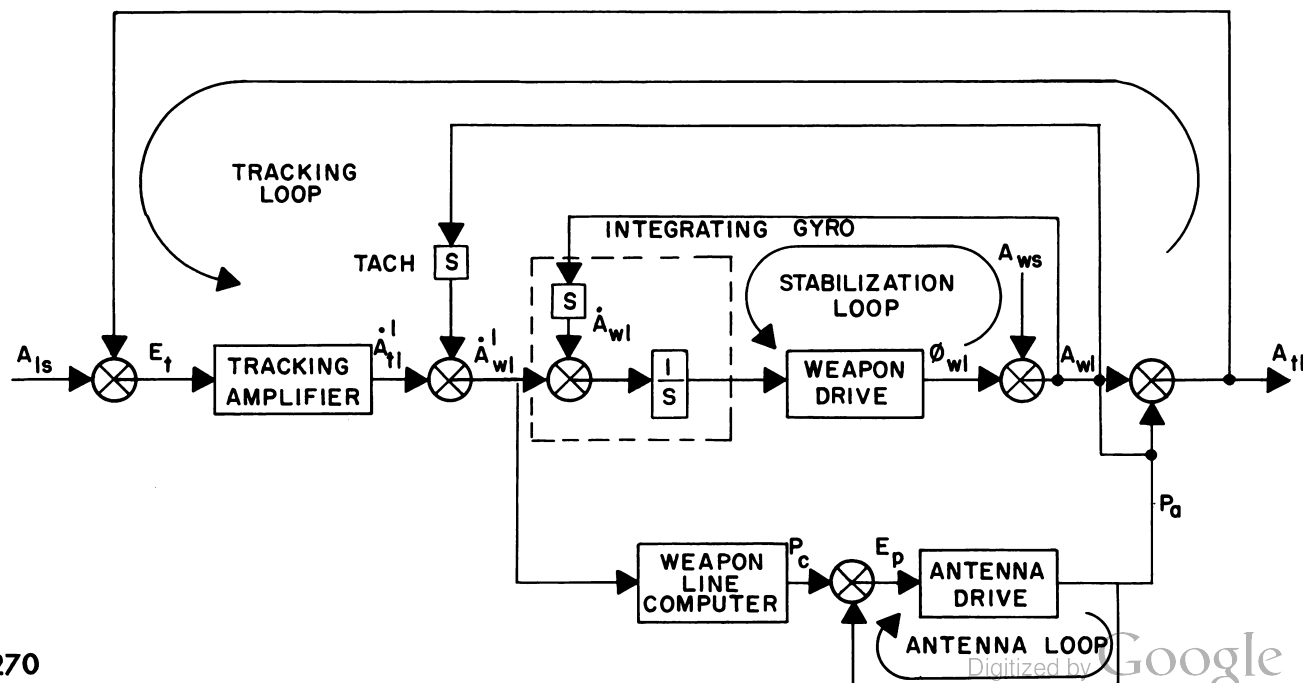
antenna-gimbal structure, and therefore the dynamic effects of gear backlash and flexure in the antenna are included within the loop. If these dynamic effects are important with respect to the weapon-drive lags, closing the stabilization loop in this manner can degrade its bandwidth. Tests on this system show, however, that the dynamic effect of flexure and backlash in the antenna is negligible with respect to other lags in the stabilization loop. In the system with the gyro, the stabilization loop does not enclose the antenna structure and the resulting configuration is not as sensitive to backlash and flexure in the antenna. On the other hand, this configuration requires that the gain of the prediction rate signal be matched with that of the weapon-line rate signal, and this matching makes system adjustment more critical.

TWO-UNIT TRACKING

A relatively new technique is being applied to weapon-control systems that is called two-unit tracking. This technique has been used in various dummy-weapon systems, systems for fixed-gun fighters, and shipboard anti-aircraft directors.

Although all the dummy-weapon systems employ weapon-line computation, it is simpler to explain the characteristics of two-unit tracking by first examining a system with tracking-line computation. An open-loop system with two-unit tracking and tracking-line computation is illustrated. The block diagram differs from that of the simpler open-loop system in the tracking section, which is illustrated here by heavy lines. The weapon-loop and computer sections of the two open-loop systems are the same.

In two-unit tracking, the operation of tracking employs two separate gimbal systems working together, one carrying the antenna and the other carrying the gyros. The radar supplies the tracking signal which is amplified and fed into an integrating gyro on the gyro platform. The signal from the integrating gyro is fed to the platform drive to close a stabilization loop about the gyro platform. To close the tracking loop, a position loop is closed about the antenna drive which slaves the antenna to follow the gyro platform. Thus,



if the position loop has sufficient bandwidth, it has the effect of a rigid mechanical coupling between the antenna and the gyro platform, so that the system operates as if the gyros were mounted directly on the antenna. Designated as A_{dtl} , the angle of the gyro platform is called the dummy tracking-line angle, because it is essentially equal to the tracking-line angle. The relative angle of the gyro platform with respect to the weapon station is designated ϕ_{dtl} .

The computer measures the torque motor input signal of the integrating gyro, which represents the desired dummy tracking-line angular-velocity A_{dtl} ; hence, it is a measure of the actual tracking-line velocity A_{tl} . The computer prediction angle P_c is added to the relative angle of the platform ϕ_{dtl} to obtain the desired weapon-line angle ϕ_{wl} , which is fed to the weapon loop. The main advantage of two-unit tracking is that the antenna can be significantly smaller since it does not carry the gyros. Since space limitations are rigid in high-speed aircraft and guided missiles, a small antenna can be quite important. The gyro platform generally can be located well within the fuselage of the aircraft, so that space limitations on the allowable size of the gyro platform are generally not as severe as are those on the allowable size of the antenna. Consequently, the two-unit tracking configuration has the additional advantage that it can more easily achieve roll stabilization of tracking and prediction, because the gyro platform often can be built with three gimbals and can be roll stabilized.

These advantages must be weighed against some important disadvantages. The two-unit tracking configuration requires two high-accuracy gimbal systems for tracking rather than one. This requirement increases complexity and tends to decrease data-transmission accuracy; the configuration may limit the tracking-loop bandwidth that is achievable with adequate stability, because the tracking loop encloses two separate servo loops. However, the latter problem is often not important because the tracking-loop bandwidth is limited anyway by sensor noise.

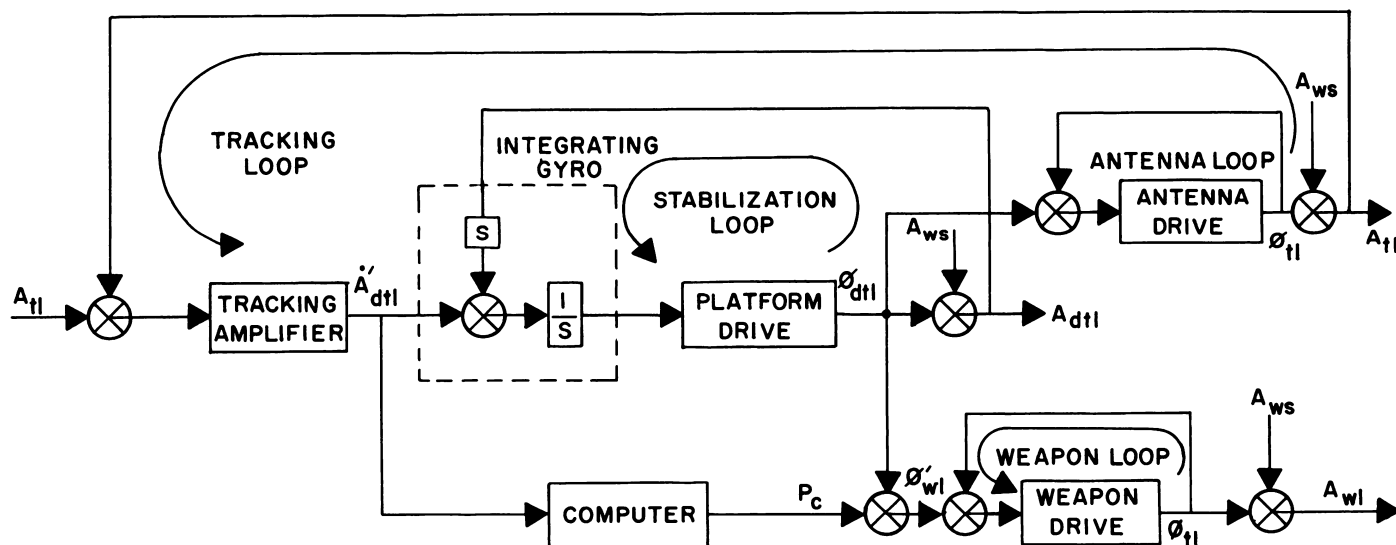
Although it may be easier to provide direct roll stabilization when two-unit rather than one-unit tracking is

employed, one should recognize that very effective indirect means of achieving roll stabilization are available. These means in general can be incorporated readily into a one-unit tracking design without requiring a three-axis gimbal system on the antenna.

Two-unit tracking can be used in a closed-loop antenna-drive tracking configuration as well as in an open-loop configuration. However, for airborne applications, use of the closed-loop configuration is generally limited to fixed-weapon systems, because a weapon turret would have to carry two gimbal systems - - the antenna and gyro platform.

The effect on the block diagram of the basic two-unit system of mounting the two-unit tracking members (antenna and gyro platform) on the weapon carriage is that the weapon-line angle A_{wl} is applied from the weapon loop down into the stabilization loop and into the tracking loop in place of the weapon-station inputs A_{ws} . The basic stability requirement is the same as for regular antenna-drive tracking systems: the stabilization loop must be much faster than the weapon loop and the prediction computer.

Actually, the term antenna-drive tracking should be generalized somewhat to include properly two-unit tracking, because the antenna drive performs only part of the act of tracking when two-unit tracking is employed. Nevertheless, the closed loop application of two-unit tracking no doubt should be in the same category as regular antenna-drive tracking systems, rather than in the weapon-drive tracking category. Although there are two units involved in tracking, both of these are auxiliary members mounted on the weapon carriage, so the tracking-line stabilization is in no way limited by the weapon-drive dynamics. The necessary generalization could be achieved by changing the term antenna-drive tracking to the more cumbersome term tracking-member drive tracking. For two-unit tracking systems, the tracking member represents not just the antenna but the two units, antenna and gyro platform, working together. Rather than making a change in terminology, however, it is more convenient to retain the original term, but realize that the antenna drive should be interpreted in a broad sense.



An open-loop, two-unit tracking system that employs weapon-line computation is illustrated. When weapon-line computation is used with two-unit tracking, the prediction angle is subtracted from the signal being fed from the platform to the antenna loop rather than being added to the signal being fed from the platform to the weapon loop. The effect of the use of weapon-line computation is to displace the platform line from the tracking line by the negative of the prediction angle, so that the platform points in the direction of the desired weapon line. Hence, the platform line in this case is called the dummy weapon line and its angle is designated A_{dwl} . The relative angle of the platform ϕ_{dwl} is fed directly to the weapon loop, so that the weapon is slaved to follow the gyro platform. Ballistic prediction components sometimes are added in the data link between the platform and the weapon, and the platform angle then

is called the future line of sight.

The torque motor signal from the integrating gyro on the gyro platform is fed to the computer, which employs weapon-line computation because this gyro signal is a measure of the desired weapon-line rate. Since the prediction signal from the computer is inserted in the tracking loop, the prediction operation tends to make the tracking loop unstable. Consequently, a tachometer is placed on the computer shaft and supplies a rate-of-prediction signal that is fed to the input of the stabilization loop, as shown by the dashed line in the illustration. If the gain in the tachometer is adjusted properly, this rate signal dynamically counteracts the prediction signal fed between the gyro platform and the antenna, so that the prediction does not disturb the tracking line and stability is achieved.

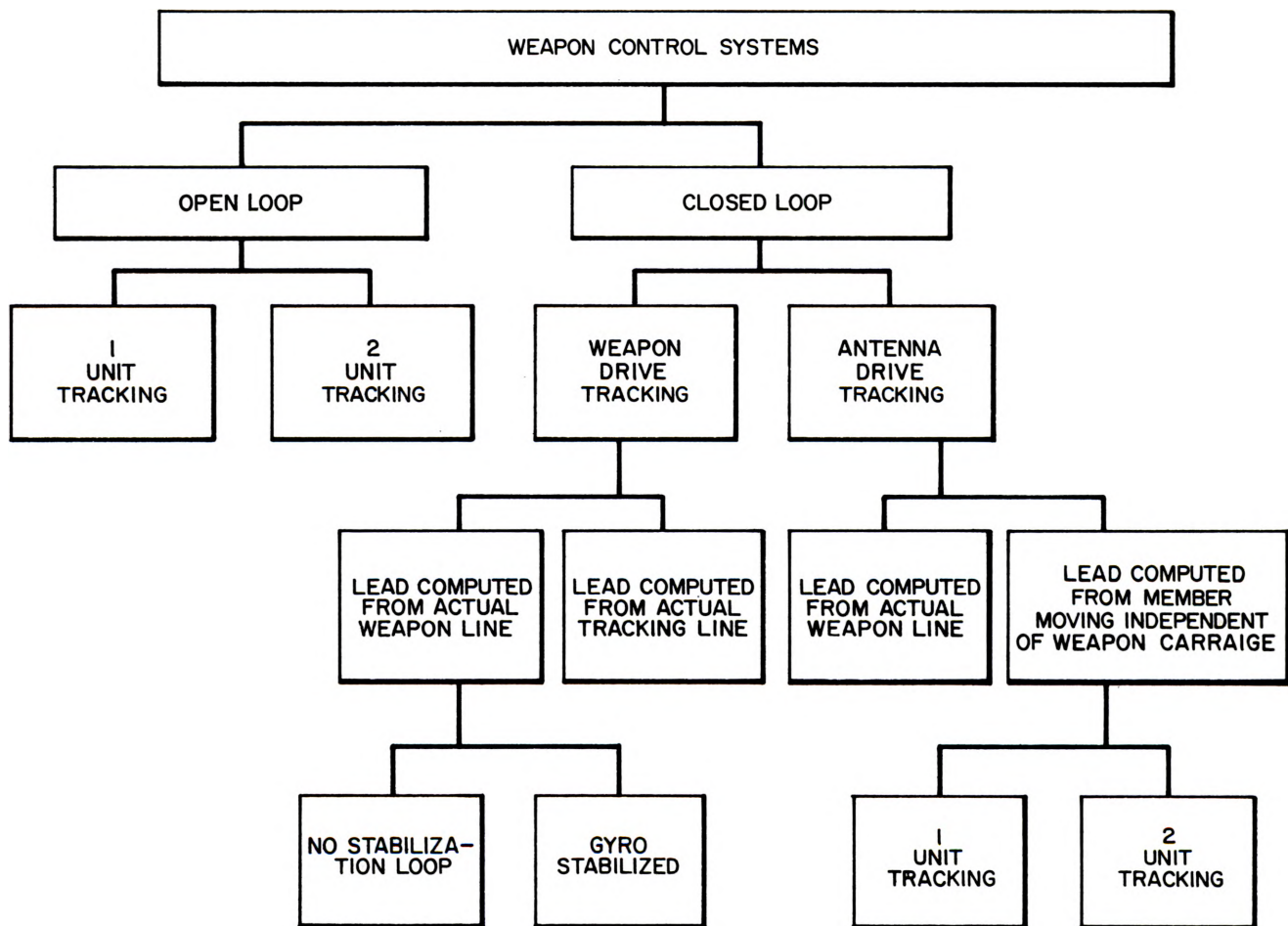
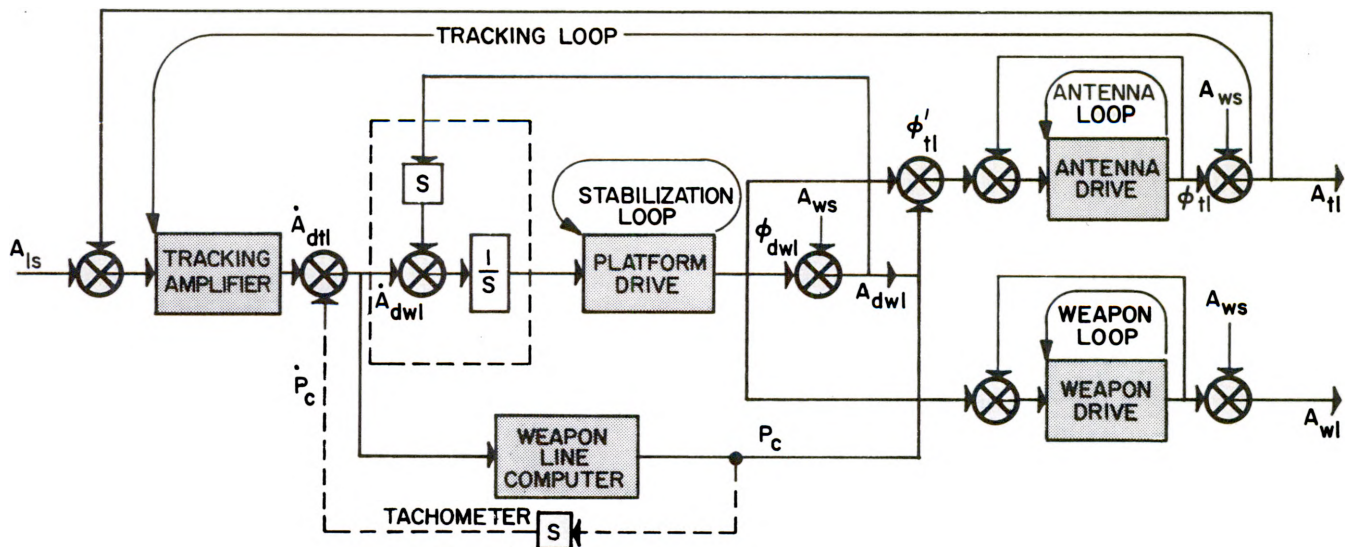
classification

The complete classification of weapon control system configurations is noted in the accompanying illustration. As has been shown, weapon-control configurations are separated into two major categories: closed-loop and open-loop. Closed-loop systems are divided into weapon-drive tracking and antenna-drive tracking. There is no corresponding division for open-loop systems, because the weapon-drive obviously cannot perform tracking unless the tracking member is physically mounted on the weapon carriage. In other words, open-loop systems must perform antenna-drive tracking.

closed loop vs open loop construction

The most important consideration in configuration design is the choice between closed-loop and open-loop construction. Open-loop construction results in simpler dynamics and generally in a smaller launcher platform. In many applications such as turrets for supersonic aircraft, space limitations are so great that a closed-loop configuration is impossible. On the other hand, closed-loop design has a decided static-accuracy advantage. For present-day high-accuracy weapon-control systems,

it is often very difficult in open-loop systems to keep mechanical compliance and backlash between the weapon carriage and the antenna sufficiently small. In addition, data transmission in a closed-loop system is inherently more accurate, because only the relative angles between the weapon line and the tracking line must be handled. These static-accuracy advantages of a closed-loop design, however, are paid for by an increase in turret size and a tendency toward inferior dynamic performance.



antenna-drive vs weapon drive tracking

Mounting the antenna on the weapon carriage in a closed-loop design can produce instability because some of the system operations tend to disturb the tracking line. The basic stability requirement for all closed-loop systems is that there be a control loop stabilizing the tracking line on the target which is sufficiently fast to correct for the disturbing effects of other operations performed within the system. The important dynamic difference between weapon-drive tracking and antenna-drive tracking is that this control loop around the tracking line is driven by the weapon drive in the former and by the antenna drive in the latter. Thus, antenna-drive tracking systems require a fast antenna drive for adequate dynamic performance, whereas weapon-drive tracking systems require a fast weapon drive. Which of these configurations is best for a given application consequently depends upon the relative dynamic capabilities of the antenna drive and the weapon drive.

The choice between weapon-drive tracking and antenna-drive tracking should largely be based upon the characteristics of the relative components to be employed. If the weapon drive is capable of a wide bandwidth with respect to the antenna drive, then weapon-drive tracking is probably the best; but antenna-drive tracking should generally be employed if the antenna drive can have a much wider bandwidth than the weapon drive.

The configurations are next separated to indicate which member supplies the measure of target rate used for lead computation. Since the bandwidth of the control loop around the weapon is inherently limited because of the large mass and size of the weapon, prediction response may have to be degraded if the actual weapon-line rate is used for computation. Thus, the chart shows a basic distinction between computing from the actual weapon-line rate and computing from the rate of a member moving independently of the weapon carriage. The independent member may be either the antenna or an auxiliary gyro platform. Since the bandwidth of the control loop of any member moving independently of the weapon carriage is not of necessity limited by the dynamics of the weapon drive, it may be possible to achieve a faster prediction response by measuring the rate of that member than by measuring the rate of the actual weapon line.

It should be recognized that the distinction between weapon-line computation and tracking-line computation, per se, is not of fundamental dynamic importance. The real consideration is not the computational technique, but the

dynamic characteristics of the member from which the computer signal is measured.

For weapon-drive tracking systems, there is no inherent dynamic advantage in computing from the tracking line instead of the weapon line, because the weapon drive performs both operations of tracking and moving the weapon. A tracking-line rate signal is just as limited by the weapon-drive dynamics as is a weapon-line rate signal. In antenna-drive tracking systems, however, we have seen that the control loop around the tracking line must be much faster than the control loop around the weapon line; therefore, the tracking line should be used for computation. If computation were to be made from the actual line in this case, a significant degradation of prediction performance might result. Similarly, for open-loop systems, prediction response might be severely limited if the lead angle were computed from the rate of the actual weapon line.

The further level of classification, which separates the configurations in terms of space-stabilization technique, actually represents two levels. First, there is the distinction between one-unit tracking and two-unit tracking. (This distinction applies only to antenna-drive tracking systems and open-loop systems, because weapon-drive tracking systems all employ one-unit tracking.) Second, for one-unit tracking systems, there may or may not be gyro stabilization. For land-based open loop systems, for example, there is no need for it. To simplify the form of the chart, however, this distinction between stabilized and nonstabilized systems is made only for the weapon drive-tracking category.

For fixed-gun fighters, antenna-drive tracking in general has a decided advantage over weapon-drive tracking, because the weapon carriage is the aircraft, which is severely limited dynamically. The control loop around the aircraft generally has a gain-crossover frequency below 1 cycle. This severely limits the allowable prediction response if weapon-drive tracking is employed, but eases the stability problem of the antenna-drive tracking configuration, because the weapon-loop bandwidth is so low that a relatively low bandwidth is adequate for the stabilization loop around the antenna. Generally, all radar-tracking fire-control systems for fixed-gun fighters employ antenna-drive tracking. On the other hand, certain optical-tracking fire-control systems for fixed-gun fighters must employ weapon-drive tracking for the simple reason that optical antenna-drive tracking could not be instrumented without dividing between two men the tasks of tracking and moving the weapon, or letting an autopilot, connected to the weapon-control system, fly the aircraft.

tracking line vs weapon line computation

As has been shown, the important dynamic distinction between tracking-line and weapon-line computation is not the computational technique, but rather whether the computer input signal is obtained from the tracking line, from the weapon line, or from some other member.

For weapon-drive tracking systems, computing from the weapon line is dynamically as good as computing from the tracking line, but for antenna-drive tracking systems and open-loop systems, prediction response may be severely limited if the computer input is obtained from the weapon.

In weapon-drive tracking systems, weapon-line computation has often been used because it has been more convenient to instrument. In some systems, for example, it would be very difficult to measure directly the angular velocity of the optical tracking line. Another example is the closed-loop weapon control system, described earlier in this chapter, in which the gyro-shaft computers were part of a computer head that was too large to be mounted on the antenna, and hence had to be mounted on the weapon line. Where tracking-line computation can be instrumented, it generally can be used in weapon-drive tracking applications, and may result in a material decrease in computer complexity.

Another application in which a component advantage was achieved at the expense of a dynamic disadvantage was the Mark 15 Gunsight for shipboard antiaircraft. This is an example of a dummy gun-line director, under open-loop, 1-unit tracking. The Mark 15 employed a modified lead-computing gunsight to control an antiaircraft gun that was too large to be driven manually. The gunsight was mounted on a director called the dummy gun and a servo loop was closed around the actual gun to slave it to follow the dummy gun. The tracking device operated as a closed loop with respect to the dummy gun, and consequently the prediction signal disturbed the tracking line. Since the tracking member was not mounted on the actual weapon carriage, the system did not have the fundamental accuracy advantage of closed-loop construction.

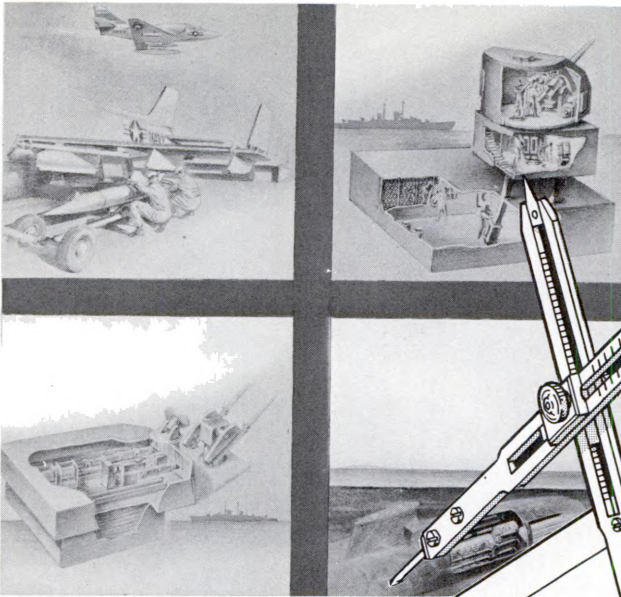
Thus, the system had some of the dynamic disadvantages of closed-loop construction without the corresponding static-accuracy advantages. On the other hand, the Mark 15 had the decided advantage of being able to employ in a convenient manner a lead computation which developed an accurate lead angle and summed it optically with the tracking-line angle in a simple yet accurate manner.

Thus, even though the distinction between weapon-line and tracking-line computation does not represent a basic means for classifying system configurations, in certain applications decided practical advantages may be achieved by computing from the weapon line. On the other hand, to use the actual weapon-line rate for computation may sometimes greatly hinder prediction response. The general dynamic requirement is that the member from which the computer input signal is measured be well stabilized against the disturbing action of the prediction signal. In other words, if the prediction signal acts as a disturbance to the control loop closed around the member being used for the computer input, then the gain-crossover frequency of the control loop must be significantly greater than the prediction-filter bandwidth.

Thus, in considering the relative dynamic effect of computing from the weapon line, the tracking line, or from an auxiliary member in the system, the important consideration is the bandwidth of the control loop stabilizing the particular member against the prediction signal. So far as the computational technique itself is concerned, the comparison between weapon-line and tracking-line computation is quite obvious. Weapon-line computation, which is basically more complex, tends to reduce accuracy. It often requires that a special correction computation be added to maintain errors at a low level when cross roll occurs. Prediction filtering is more difficult to control with weapon-line computation. On the other hand, weapon-line computation sometimes may provide a much more convenient way of determining target angular velocity than if the tracking-line rate is measured directly. It sometimes makes a simple yet accurate summing of the prediction angle with the tracking-line angle possible. This can result in a net simplification of the overall computation and resolving process.

one-unit vs two-unit tracking

Two-unit tracking may be employed both in antenna-drive tracking systems and in open-loop systems. The choice between one-unit tracking and two-unit tracking is fundamentally a compromise between antenna size on the one hand and system complexity on the other. Two-unit tracking allows the size of the antenna to be decreased, but requires a significant increase in system complexity.

**WEAPONS
SYSTEM****design
and
development**

Throughout the course of history, the development of weapons has been extremely slow. Until 1000 A. D., weapons had a useful life of about 400 years. Today, the life span of a weapon system is from 3 to 7 years. Development of a modern weapons system requires considerable time. Weapons systems therefore tend toward obsolescence almost before the first production units enter the field. The design and development phase

has become increasingly important as the useful life span of weapons has decreased, because it is the speed of development which determines the extent to which the system design incorporates the latest technical advances. This chapter is not intended to cover design and development techniques, but only to demonstrate the interface between development and utilization and to give some insight into the nature of the task of systems engineering.

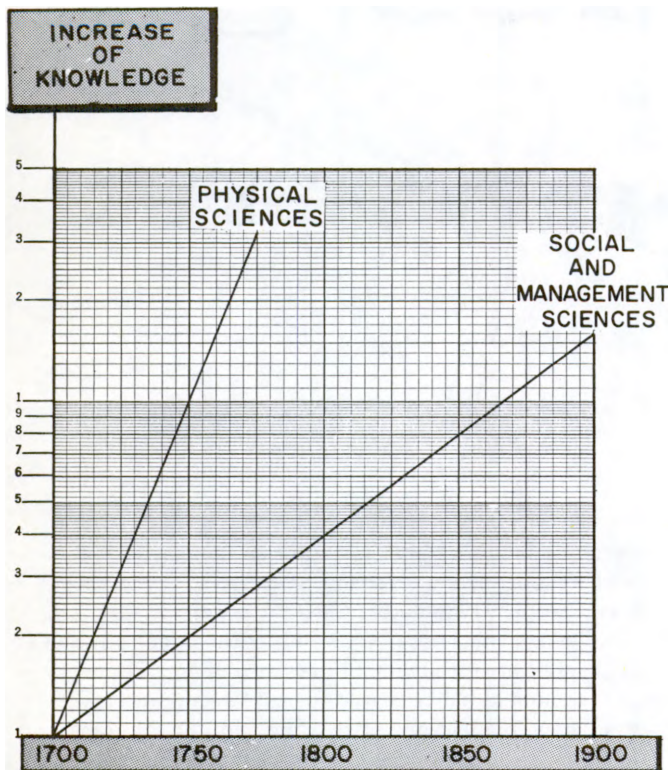
Imagination in weapon system design

Revolutionary weapons come from new ideas, innovations, and technical breakthroughs resulting from the creativity of man's mind, not from modification of previous weapons. When new weapon ideas are conceived their full potential must be realized. The longbow, for example, was available and advocated as an important weapon in the 12th century. It was not until 250 years later at the battle of Crecy, however, that the longbow achieved its position of prominence.

Such failures to recognize weapon potentials are not confined to the past. In 1940, when the Germans were reputedly doing vast research for the atomic bomb, only \$6,000 was available in the federal budget for atomic studies. The prompt recognition of the importance of new ideas and technological breakthroughs and of their military significance is essential to maintain an effective deterrent force.

effects of increase in knowledge

It is reasonably evident that knowledge in the physical sciences has been increasing at an exponential rate since the end of the seventeenth century and is swiftly approaching that position of the exponential growth curve which has an infinite slope. While knowledge in the physical sciences is doubling every 15 years, the ratio of increase is much slower in the social and management sciences, doubling approximately every 50 years.



Two of the major effects of scientific growth on weapon development are increased cost and complexity. If it is assumed that the number of choices in the design of a weapons system is proportional to the amount of existing physical knowledge, then we can readily see that the current weapons system designer has a tremendously greater number of alternatives in design approach than his predecessor. It is also evident, since the amount of knowledge is increasing exponentially, that the number of design alternatives is also increasing exponentially.

Since the increase and rate of increase of knowledge in the physical sciences has made the future so uncertain, weapon systems must be planned much further ahead and must be flexible enough to accommodate unexpected changes in situations and technologies. It is necessary to estimate well ahead of time the costs and values of doing things in different ways so that when the time comes to make a decision, the problem will already have been analyzed. In this manner, the proper decisions may be made without costly wastes of time and with greater confidence.

To find the proper approach to a weapon system design, a careful survey of all the alternatives must be made. Overlooking the best methods available will result in a less than optimum weapon system as well as a waste of the technical effort and knowledge expended to develop the better methods.

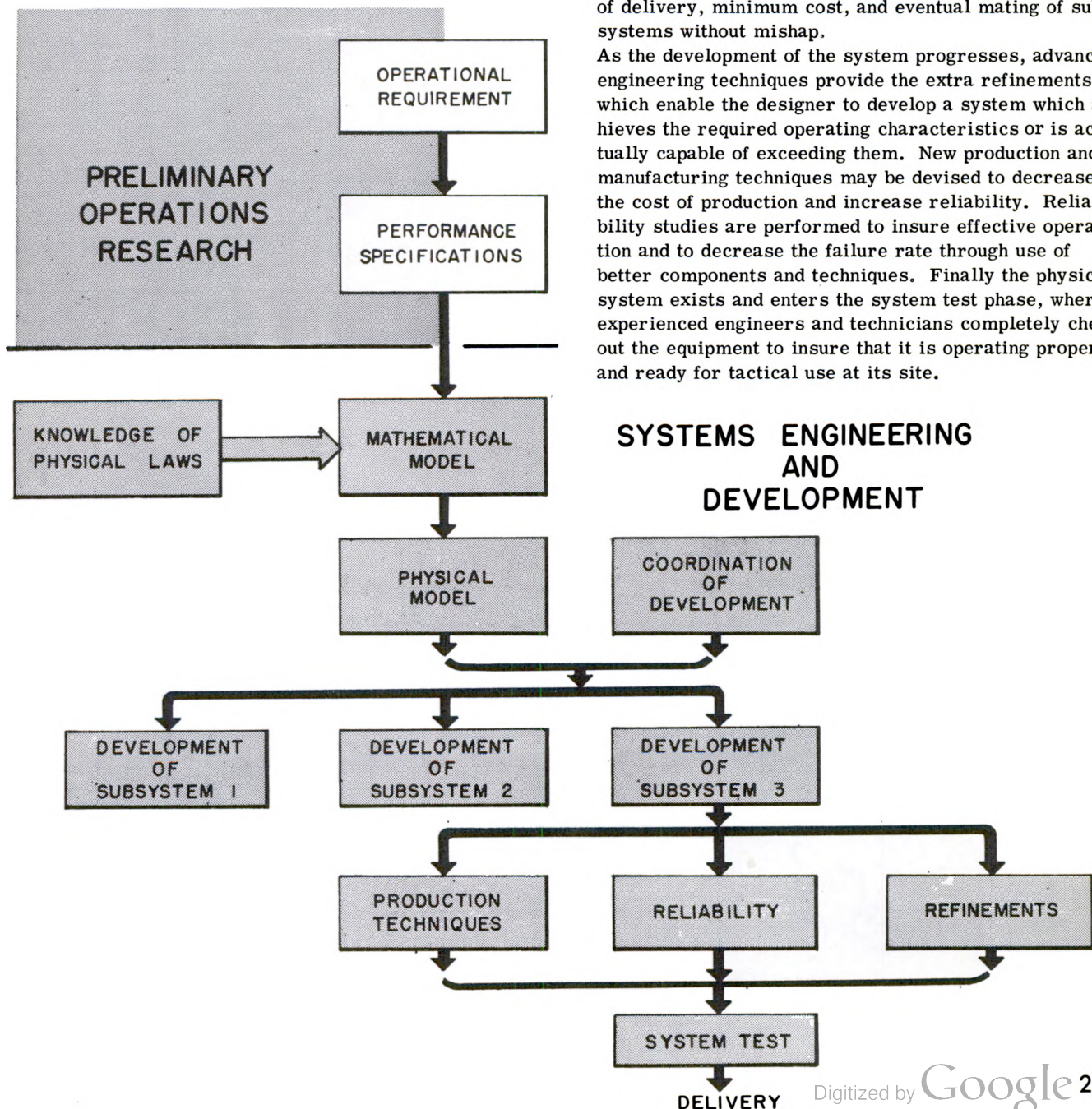
weapon considerations

In general, the best weapon is the most economical one to design, build, and use which will effectively achieve the desired military objective. Since the weapon system designer is faced with an ever increasing number of alternative design approaches, it is extremely important that he choose those alternatives which will most economically produce a tactically reliable weapon. Low-cost weapons have the added advantage of being producible in large numbers. Thus more of them are available for a given sum of money. In general, a low-cost weapon is also a simple weapon. Simplicity of design facilitates weapon operation and consequently increases reliability. It is also easier to train personnel effectively to employ and maintain simple weapons, thereby minimizing the amount of time necessary for effective mating of weapons and manpower.

compatibility of existing military doctrine with new weapon systems

Military strategy and doctrine must be kept abreast of weapons system developments. Work on revising doctrine brought about by the development of new weapon systems must begin with the conception of the system so that the new system can be used effectively by the fleet immediately upon issue.

The evolution of a weapon system begins with the formulation of new ideas and the existence or forecast of fleet requirements for such a system. The weapon system approach to advanced planning is used to initiate and realize a new system based on the predicted requirements of the fleet. It is customary to develop a complete weapon system, including missiles, below-deck equipment, support equipment, personnel, and operating plan. The Polaris missile system is a prime example of the use of advanced weapon system planning.



the weapon system approach to design

Operational requirements and performance specifications for a weapon system are determined by operations research. Once the requirements of a weapon system are established, a mathematical model of the system is built. The mathematical model is necessary to evaluate the technical feasibility of the proposed system. Once the mathematical model has proven the proposed system in theory, a physical model is built. These physical models may consist of wind tunnel models, breadboards of particular circuits, and samples of critical and novel components.

The actual development of the system is divided into a parallel development of various subsystems. The coordination of these various subsystem programs requires active exercise of the most advanced management techniques. Report and evaluation programs assisted by computers are employed to insure timeliness of delivery, minimum cost, and eventual mating of subsystems without mishap.

As the development of the system progresses, advanced engineering techniques provide the extra refinements which enable the designer to develop a system which achieves the required operating characteristics or is actually capable of exceeding them. New production and manufacturing techniques may be devised to decrease the cost of production and increase reliability. Reliability studies are performed to insure effective operation and to decrease the failure rate through use of better components and techniques. Finally the physical system exists and enters the system test phase, where experienced engineers and technicians completely check out the equipment to insure that it is operating properly and ready for tactical use at its site.

SYSTEMS ENGINEERING AND DEVELOPMENT

SYSTEM DESIGN

effectiveness

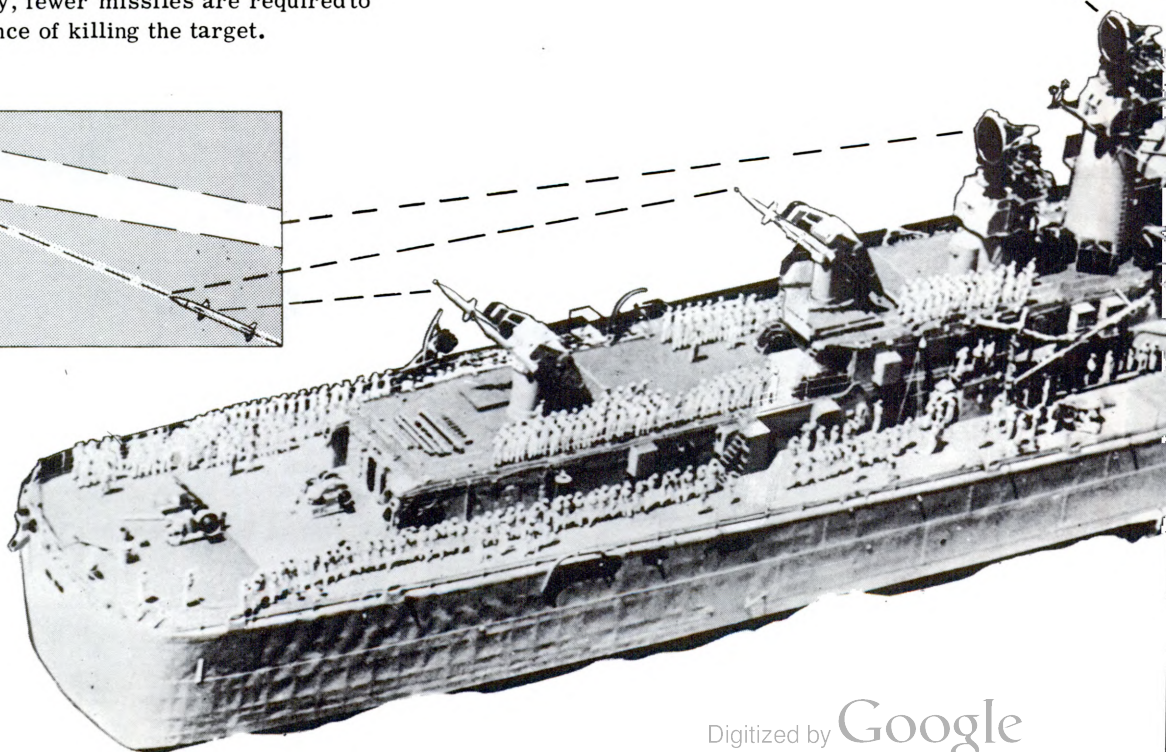
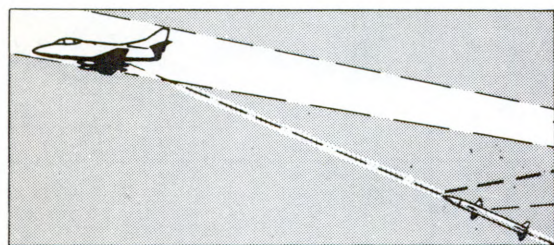
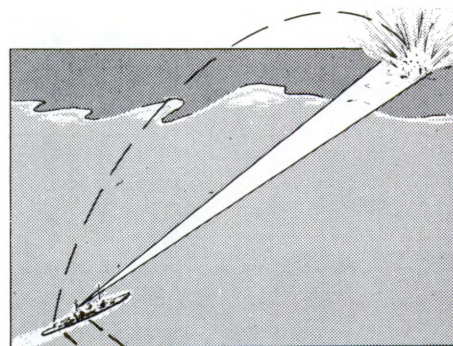
The end objective in the development of any weapon system is to achieve maximum damage to the enemy with the least cost to ourselves. Everything may be considered to have a cost in terms of dollars. The cost of equipment may be found by determining what it would cost to manufacture a replacement. The fiscal value of a human being may be determined by calculating the cost of recruiting, educating and training his replacement. Such things as morale cannot be evaluated easily. However, their value may be calculated because the production loss or production slowdown caused by their absence can be estimated in monetary terms. In a similar fashion, the cost of the weapon system and the missile can be determined. By dividing the value of the damage caused by the total cost of the weapon system, we have a number which is an indication of the effectiveness. From this it can be seen that the effectiveness of a weapon system can be increased by increasing the destructive capabilities of the missile, by decreasing the cost of the missile, or by a combination of both.

kill probability

A class of missiles which always destroys its target has not yet been developed. When the relative merits of a weapon system are discussed, one of the most important factors is missile probability of kill at the target. A low missile kill probability requires that a greater number of missiles be launched. An effective weapon system can still be achieved if the cost of the individual missiles is low and a high rate of fire can be achieved. A machine gun is an example of a weapon system in which high rate of fire is combined with low cost of individual rounds. Although individual missiles with high kill probability are more costly, fewer missiles are required to attain the same assurance of killing the target.

performance

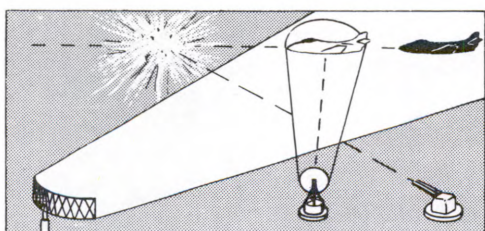
The effectiveness of the weapon system is necessarily dependent in part on the performance characteristics of the system. Performance may be described in terms of such characteristics as accuracy, speed, stability, range, and target capacity. A missile which meets minimum requirements for each of these characteristics and still achieves a high kill probability for a particular mission may be lower in cost per individual weapon and may thereby achieve higher system effectiveness than a missile which meets higher requirements for individual characteristics.



REQUIREMENTS

operability

Simplicity of operation enhances the usefulness of a system. Easily read control panels, simple operating procedures, and so on increase the efficiency of the system. Furthermore, since there is no segment of the globe which is not a potential battleground, the ability to operate in all kinds of weather and environment is a highly desirable quality of a weapon system, from both a military and financial point of view.

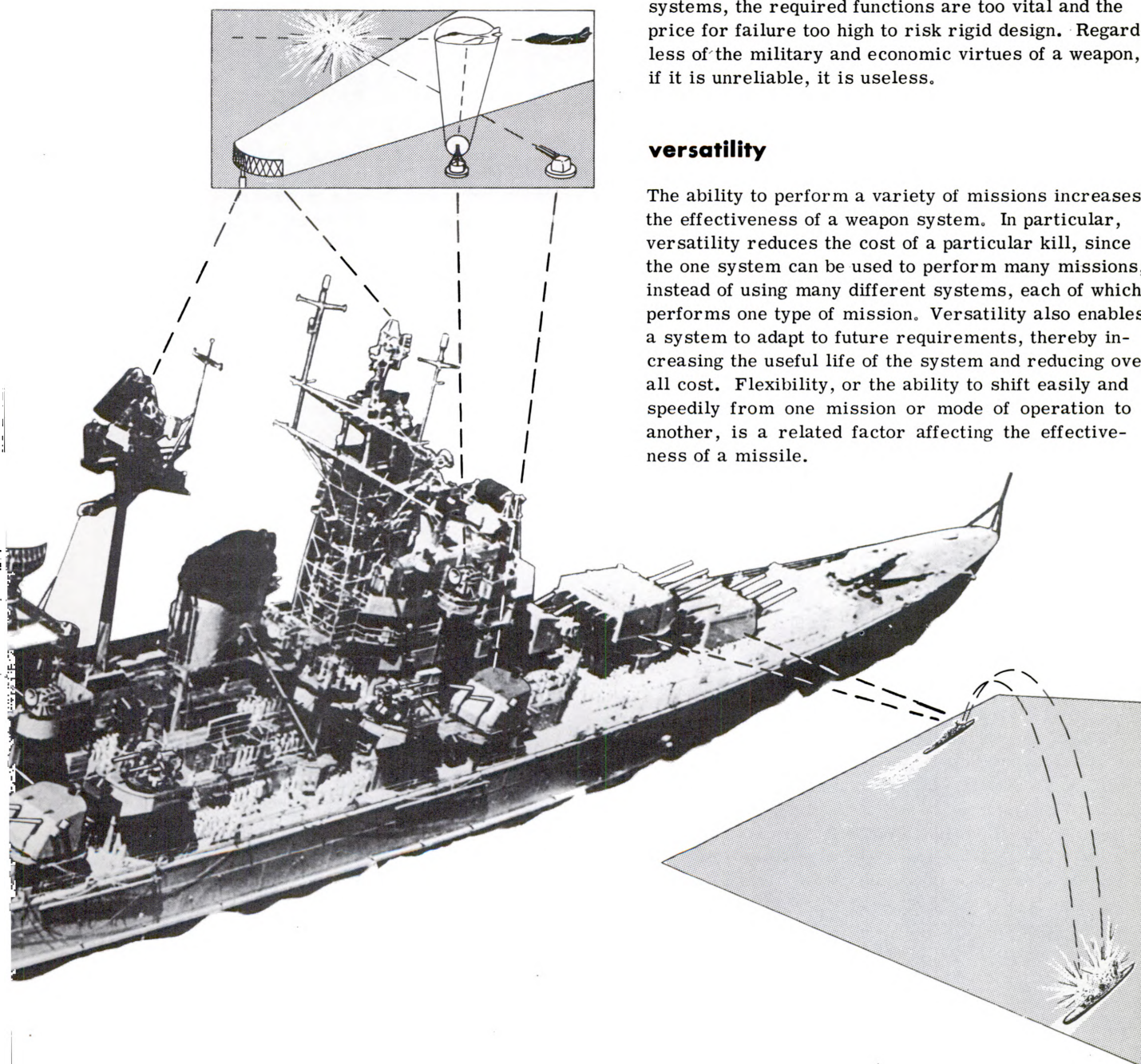


reliability

Reliability is so important in weapon systems that it may be emphasized even at the expense of other factors contributing to an effective system. System reliability has been defined as the probability that the system will perform its required function under given conditions for a specified operating time. Efficiency should not be increased at the expense of reliability. While rigid design will produce an efficient, highly reliable system in the laboratory, the use of tight tolerances and safety factors in the field may result in system failure. In weapon systems, the required functions are too vital and the price for failure too high to risk rigid design. Regardless of the military and economic virtues of a weapon, if it is unreliable, it is useless.

versatility

The ability to perform a variety of missions increases the effectiveness of a weapon system. In particular, versatility reduces the cost of a particular kill, since the one system can be used to perform many missions, instead of using many different systems, each of which performs one type of mission. Versatility also enables a system to adapt to future requirements, thereby increasing the useful life of the system and reducing overall cost. Flexibility, or the ability to shift easily and speedily from one mission or mode of operation to another, is a related factor affecting the effectiveness of a missile.



Invulnerability

Invulnerability is the ability of a weapon system to resist defeat or damage by the enemy. The weapon system must be designed to be as invulnerable as possible. Maximum security against enemy action and environmental conditions must be provided. Enemy action in the form of countermeasures can be combatted by establishing an active system, a counter-countermeasure, to counteract the enemy action or by strengthening the weak links in the weapon system against which the enemy countermeasure is directed. Certain weapon system components, such as radar antennas, particularly above-deck components, are highly susceptible to battle damage. Many of these components can be strengthened to resist damage, while others can be set up for easy repair or the quick use of alternate, undamaged components.

Maintainability

System maintainability must be designed into the weapon system so that maintenance may be effectively performed by shipboard personnel. The required degree of maintainability is determined by the weapon systems mission, location, environment and maintenance personnel. For instance, a particular weapon system may be required to operate continuously. This system will require a minimum of down time. The weapon system designer may utilize techniques such as automatic built-in test equipment to isolate a malfunction rapidly. He may use replaceable modules, so that once the built-in equipment isolates the malfunction to a module, the maintenance technician may remove the faulty module and replace it with a spare. The module may then be repaired without interfering with overall system operation. Easily accessible, strategically placed test points may also be designed to reduce down time and enhance maintainability. Further factors which must be considered in designing a weapon system from the maintenance point of view are the storage, cost and accessibility of spare parts.

Summary

In general an effective weapon system must be able to accomplish the objectives which initiated its design in the most efficient and cheapest way possible. In planning such a system, a designer must be conscious not only of the cost and critical requirements of the proposed weapon but also of the nature and value of the target it will be employed against.

While all of the previously discussed system requirements are essential to effective weapon system design, reliability is the most important consideration, even if it has to be bought at the expense of the others. For if a weapon cannot be depended upon except under the most ideal conditions, it is rendered virtually useless.

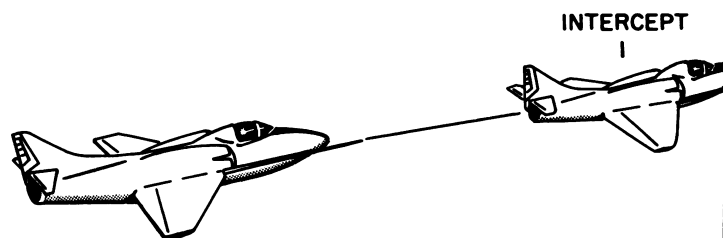
DETERMINATION

New operational requirements are generally determined by three factors: 1) disclosure of a new and potent weapon in the hands of a prospective enemy 2) realization of serious shortcomings in our existing weapons 3) recognition of the military implications of technological advancements.

new and potent weapon

The disclosure of a new and potent weapon in the hands of a potential enemy necessitates the acquisition of as much information as possible about this weapon so that its threat, performance, effectiveness, etc. may be evaluated as accurately as possible. Once this information is available and an assessment of the new weapon's effect on our military posture is made, an effective countermeasure can be designed at minimum cost and with minimum effort. There are many ways in which an increase in enemy weapon potential can be counteracted. A change in tactical defense doctrine may suffice, or a completely new doctrine may have to be devised. If existing weapons are flexible they can be altered to meet the latest requirements; if not, an entirely new weapon system will have to be designed. The selection of the best method to offset an increase in enemy weapon potential involves three basic design tasks:

- 1) Formulation of optimum tactics and/or a countermeasure device.
- 2) Demonstration of feasibility of tactic or countermeasure.
- 3) Evaluation of solution, using total damage per unit of cost as a measure of effectiveness.



OF OPERATIONAL REQUIREMENTS

Operations research very often discloses better ways to use existing weapons, thereby removing their shortcomings. For instance, if it is necessary to defend an aircraft carrier against high-altitude, high-speed bombers and existing antiaircraft weapons are known to be inadequate because of the long time of flight of the missile and the relatively high maneuverability of the attacking bombers, does a requirement for a new surface-to-air missile system exist?

serious shortcomings in existing weapons

A different problem arises when the difference in weapon capability results from lack of progress on our part, rather than from progress by the enemy. Consequently, the first consideration must be the nature of the targets that will exist at the time the new missile system becomes operational. In some cases, the development time required for a new missile system exceeds the time originally estimated, making the system inadequate against the improved targets which have been developed during this extra time.

In some cases, missile systems have been made more complex than required for the targets for which they were designed. Unless specific information is known about present and future enemy weapon systems, it is often best to assume that the enemy has the same military capability as we. Therefore for every weapon in our arsenal there should be an effective counterweapon.

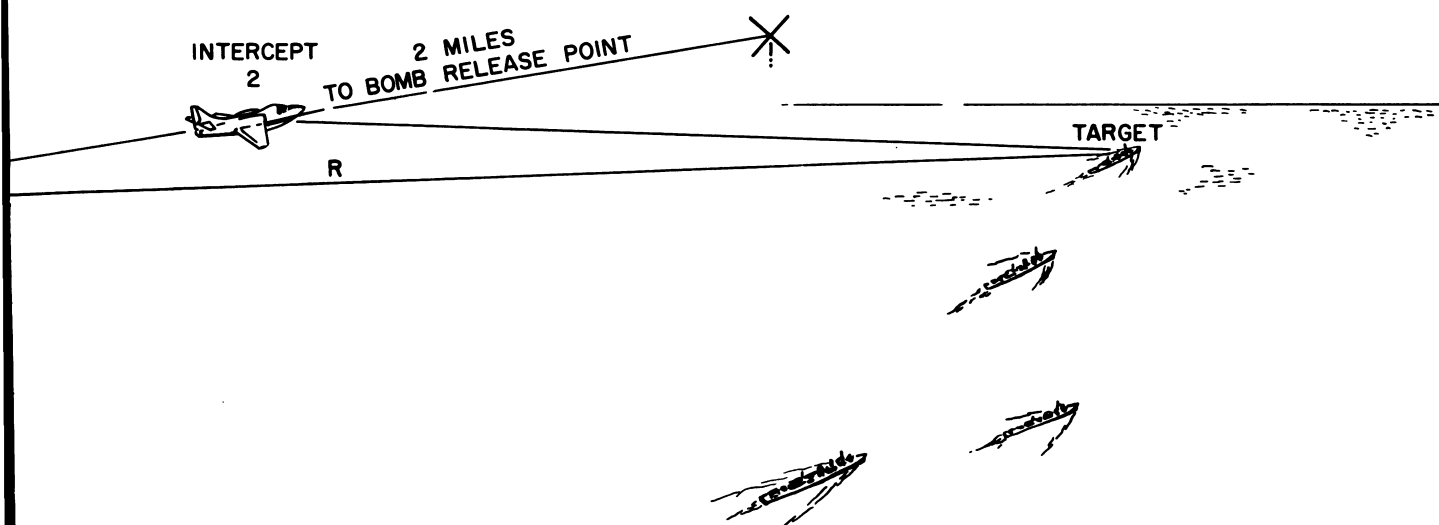
TARGET CONSIDERATIONS

Assume that target speed and maximum altitude are known. Additional pertinent data about the target to be evaluated might comprise its vulnerability, electronic countermeasures capability, maneuverability, etc. It will also be necessary to know if the weapon is to be used against a single target or against multiple targets attacking in formation or simultaneously from several directions.

DEFENSIVE SYSTEM REQUIREMENTS

After analyzing the target's capabilities and performance, it is necessary to examine the weapon system capabilities required to effectively counter the threat posed by the attacking target. If we assume that because of other methods of defense, it is necessary to defend the flight path illustrated for up to 10 attacking targets only, in formation or attacking singly, we may formulate the overall system capabilities as follows:

- 1) Must have a kill probability of at least 0.95 for each attacking target prior to bomb release.
- 2) Must achieve this kill probability for up to 10 bombers flying singly or in formation and in all weather conditions.
- 3) The early warning system must initiate defense operations in sufficient time to accomplish 1 and/or 2.



AVAILABLE DEFENSIVE SYSTEM CONSIDERATIONS

Assume that the defensive missiles available on the carrier have an average speed compatible with target speed. Beam-rider guidance is provided by a sensor system consisting of a target tracking radar and a guidance beam radar. The system is capable of guiding only one missile at a time and possesses a single-shot kill probability of 0.80. To achieve a kill probability of 0.95, it is necessary to fire two missiles at each target. Since the probability of the attacking target's surviving a single missile attack is 0.20, the probability of surviving two missile attacks is $(0.20)(0.20)$, or 0.040. Therefore, the kill probability, using two missiles, is 1 minus 0.04 or 0.96. Since the defensive system available is capable of firing and guiding only one missile at a time, it is necessary to fire two missiles successively rather than concurrently. In this manner cost can also be reduced, since a kill by the first missile would eliminate firing a second. Thus, weapon effectiveness is increased.

The sequence of events for two successive firings, employing typical time values, is as follows:

Time (minutes)	Action
0	Detect target
3	Identify target and alert crew
8	Ready missile system and acquire target
9	Assess damage from intercept and fire next missile if necessary
10	Make intercept

If the enemy deploys his craft in a simultaneous attack from separate directions, a possible method of defense would be to place launching stations about the perimeter of the point being defended. If this method is shown to be operationally and technically possible, the only remaining factor to be considered is the logistic feasibility of the method. In analyzing the logistics of a situation the following questions must be considered:

- 1) What manpower requirements in numbers and skills exist?
- 2) What existing weapon systems will be replaced and to what extent can the manpower concerned be utilized by the new system? Is superiority over these weapon systems clear?
- 3) What is the initial cost of the multisite system, including wartime missile allowances?
- 4) What training costs are involved?
- 5) To what extent are these costs offset by those of replaced weapons?
- 6) Can the national resources meet these costs and manpower requirements?

Assuming that the multiunit system suggested cannot achieve the required operational results against a mass formation target, modifications to the system, such as use of a nuclear warhead in the missile, may provide the answer.

Using these time intervals, it is possible to calculate the points along the target's flight path at which these events take place and the relationship of these points to the final intercept point, which should be at least two miles before the target's bomb release point.

The maximum range required of the defensive missile system is the distance (R) to first intercept. Using the time figures given, the available sensor or early warning system must be sensitive enough to detect the bomber ten minutes before it is within striking distance of the carrier. It must then be able to identify the target and alert the missile crew within 3 minutes. Based on this analysis, the available system meets the performance requirements necessary to defeat the attacking targets and also appears to be operationally feasible for the single target case.

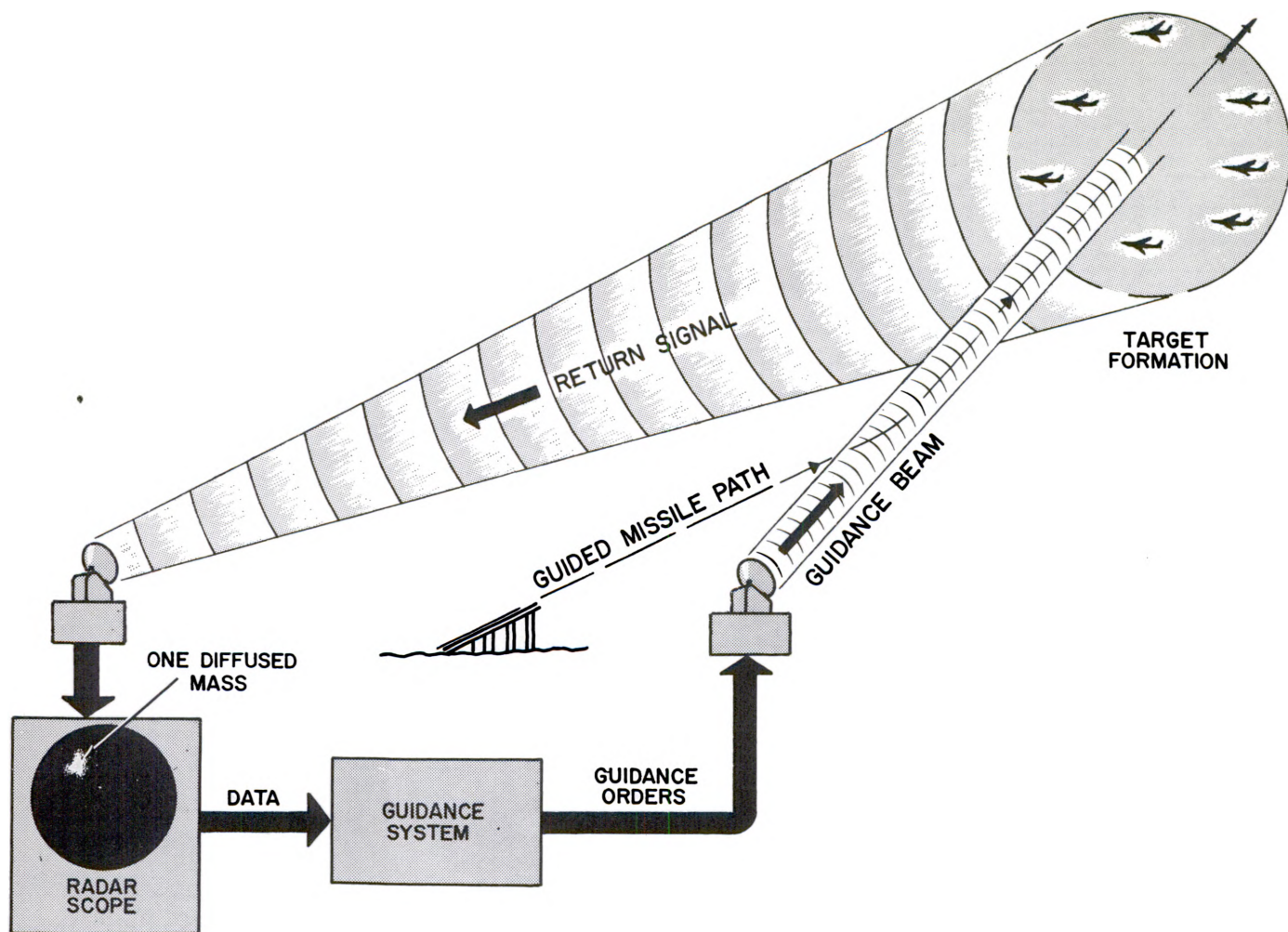
If the enemy bombers are in formation, the massed attacking force provides a much larger radar target, thereby decreasing the enemy's chances of surprising the defense. In this case, earlier and surer warning is provided to the defense force, allowing time for the carrier to send out interceptors to impose casualties on the attacking force long before it comes within range. Thus, the number of targets that must be destroyed by the carrier's missile defense system is reduced.

REASSESSMENT OF ASSUMPTIONS

After a complete analysis of the problem, the existing defense system may be found adequate or a new weapon system may have to be designed. Before making a definite decision the operation research man must make a complete review of the entire analysis, including a complete reassessment of all assumptions upon which the analysis is based.

Since much of the analysis is based on assumptions, it is particularly important to examine these very carefully. For example, the proposed missile system might be reexamined and certain parameters changed to protect against enemy actions such as:

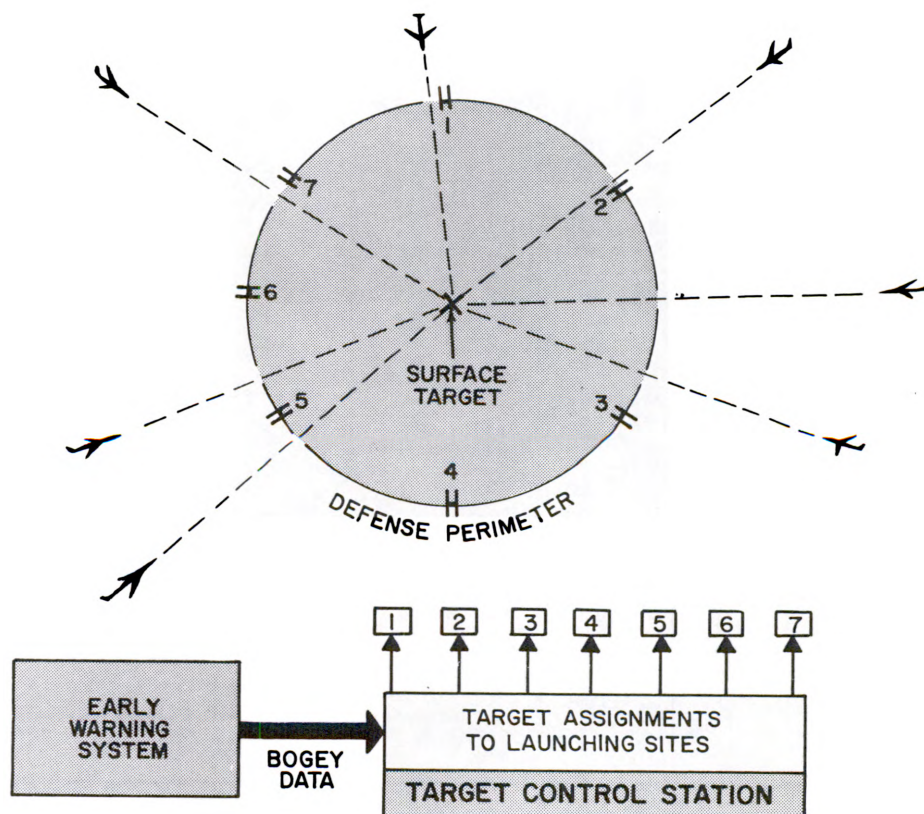
- 1) Use of low-altitude attacks.
- 2) Use of attacking craft equipped with air-to-surface



new innovations and advanced technology

New innovations or a revolutionary advance in technology may lead to the issuance of a new operational requirement. To illustrate the process of operational research used to aid in the decision whether a new operational requirement should be issued or not, consider the following problem:

There exists a present system by which nuclear warheads may be delivered against targets of naval interest by carrier aircraft. Does an operational requirement for delivery of such bombs by guided missiles exist?



PRELIMINARY ASSESSMENT

The first step in analyzing this problem is to study the relative advantages and disadvantages of guided missiles over carrier aircraft. Some of the advantages are:

- 1) Loss of a missile does not mean loss of a highly trained pilot.
- 2) Unmanned missiles need travel only one way, consequently increasing striking range for a given gross takeoff weight.
- 3) Tactics which insure personnel safety from own nuclear blast need not be employed.
- 4) Extra equipment and gear required for a pilot are not required on missiles, thereby giving a performance advantage to the missile.

Some of the disadvantages are:

- 1) High cost resulting from nonrecoverability
- 2) Tactical inflexibility and poor reliability resulting from the absence of a pilot.

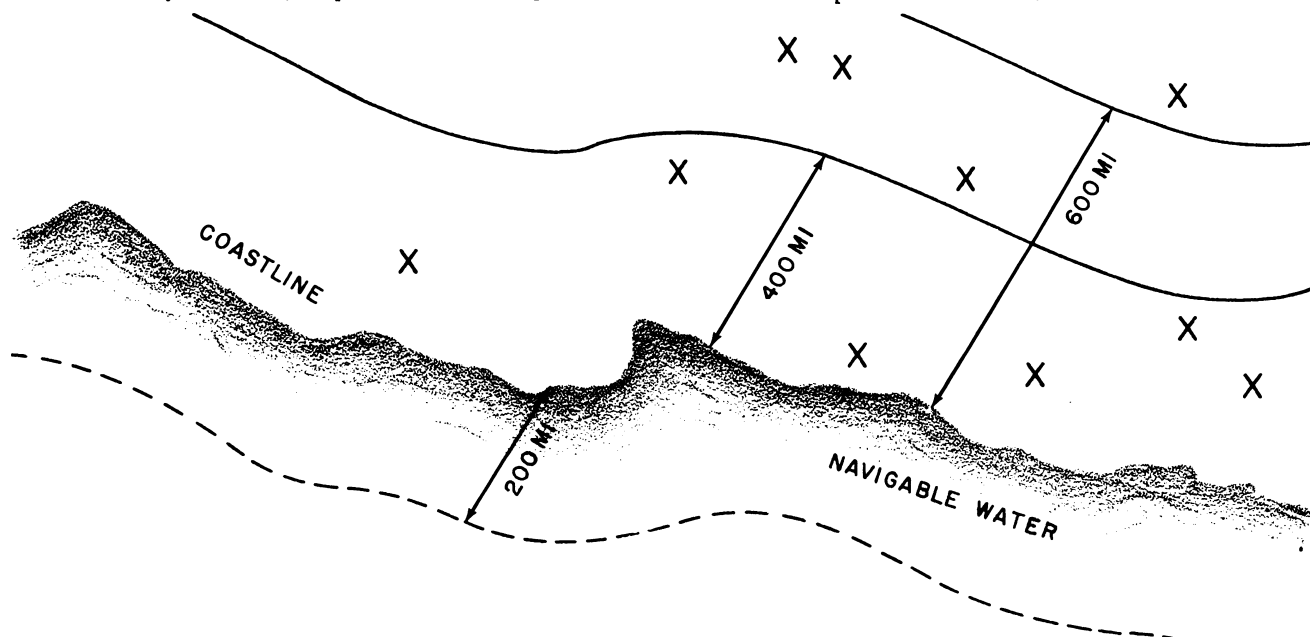
TACTICAL CONSIDERATIONS

The next step after the preliminary assessment is to analyze the tactical considerations of the problem, such as environment, target, launching vehicle, and trajectory. Since the missile is launched from a ship, the operating environment of our launching vehicle is defined as being any navigable water. In addition it is most economical to place our launching vehicle as close to the target as the situation allows. By studying the targets of the enemy in relation to navigable waters, the range required by the missile can be found. An additional range factor should be added to allow reaching these same targets should close approach be dangerous because of heavily defended coastlines. For example, assume that 90 percent of the targets lie within 600 miles of navigable water, and that 60 percent of the targets lie within 400 miles of navigable water. If a minimum of 60 percent of the targets must be available at all times, a good design range would be 600 miles. Thus, if the ship had to lay off shore 200 miles because the coastline was heavily defended, 60 percent of the targets

would still be within striking distance.

The next question which must be studied in this analysis is: what defenses must be penetrated to reach these targets? Some of the more important factors in determining defense capabilities are:

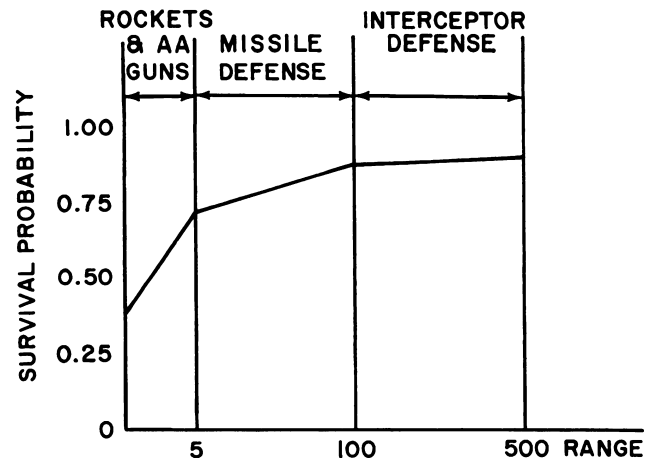
- 1) Effectiveness and kill capability of enemy interceptors, surface-to-air guided missiles, and short-range AA guns and rockets.
 - 2) Speed, altitude, and radar reflectivity of our attacking missile.
 - 3) Performance capability and disposition of the enemy's early warning network and command control system.
- When these questions are answered, computation of survival probabilities for various missile approach speeds and trajectories can be made. These probabilities may be plotted graphically to illustrate the best possibilities. In this manner, the best combination of missile performance and trajectory for defeating the enemy defense may be selected. Since the missile must survive the enemy defense in order to deliver its warhead, the survival probability of the missile is a critical measure of predicted success.



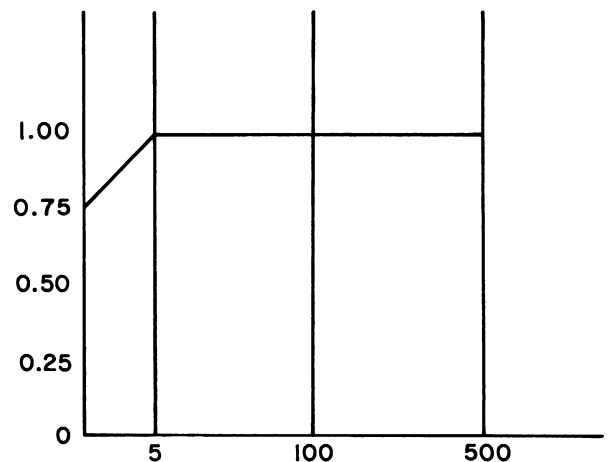
The method of determining survival probability against air defense may be roughly outlined as follows:

- 1) Determine, from wartime statistics or other means, the kill probabilities of likely enemy weapons such as interceptors, guns, and rockets.
- 2) Adjust these probabilities for estimated improvements in these weapons since the statistics were generated.
- 3) Further adjust these probabilities for the differences between our missile and the type of targets for which these statistics were valid.
- 4) Estimate the number and disposition of weapons capable of intercepting our missile and also estimate the capabilities of the enemy's early warning system, target identification system, and weapons control system.
- 5) Estimate the number of weapons which may be successively brought into action against each of our missiles and, consequently, the survival probability of our missile at the various ranges at which these weapons will be brought to bear. Then estimate the cumulative survival probability as a function of range to target. Note that the estimates made concerning enemy defenses should reflect the predicted capability of the enemy's defensive weapons at a future date to allow for development time of the proposed weapon system.

Consider the survival probability for two proposed missiles with different performance-trajectory combinations. First assume a pulse jet missile with a cruise altitude of 40,000 feet, a speed of Mach 1.0, and a final vertical dive over target. From a range of about 500 to 100 miles, the missile is engaged by enemy interceptors of slightly superior performance. Detection and tracking of the missile is facilitated by its relatively high altitude and moderate speed. Also, the inability of the missile to detect and evade an attacking interceptor adds to the effectiveness of the enemy defense. In this range band, the missile's survival probability is 0.85. From 100 to 5 miles, the missile is engaged by surface-to-air guided missiles which effectively reduce the cumulative survival probability of the missile to 0.7 at 5 miles. From here on in, rockets and AA guns further reduce the cumulative survival probability to 0.04. Therefore, we can say that an average of 4 out of every 10 missiles will survive the enemy's defense and deliver their warheads. Now, consider a second missile, also a pulse-jet, with a speed of Mach 1.0, employing a low-altitude attack with a final climb to burst altitude over the target. Because of the low altitude, the enemy's early warning system is hampered in its detection of the missile. This also handicaps effective action in directing guided surface-to-air interceptors. In going through the interceptor and guided missile defense bands, the missile has its survival probability reduced to 0.9. Now, when the missile climbs toward its burst point, AA guns and rockets impose a severe attrition, causing the final cumulative survival probability to go down to 0.7. To summarize, the first proposed missile has a survival probability of 0.4 and our second proposed missile has a survival probability of 0.7.



pulsejet missile — altitude 40,000 ft. — mach 1.0



pulsejet missile — low altitude — mach 1.0

REASSESSMENT OF ASSUMPTIONS

As explained in the analysis for shortcomings in existing weapons, it is now necessary to reexamine all the assumptions made during the analysis. For instance, if a large surface vessel were employed as the launching vehicle, it would be possible to fire a number of missiles in succession, thereby saturating the defense. However, this would also cause loss of surprise and a channeling of missile trajectories would occur. By choosing inclement weather, the loss in survival probability due to interceptors might be decreased.

It is also necessary to correct the survival probabilities for the two missiles to account for the fact that these missiles do not have a 100-percent reliability factor. For example, if 90 percent is the reliability factor for the first missile, its survival probability is decreased from 0.4 to 0.36. Also, if the second missile has an 80-percent reliability factor, its survival probability is decreased from 0.7 to 0.56.

As a result of reassessment, the final figure of merit for each missile which takes reliability into account in terms of survival probability is as follows:

- First missile - 0.36 survival probability
- Second missile - 0.56 survival probability

TECHNICAL FEASIBILITY

It is now necessary to determine the technical feasibility of both missiles, including the extent of shipboard alterations required, and the degree to which both missiles and aircraft can be operated from the same ship. Assuming that both missiles are feasible, the next question is: How long will it take to develop each missile? While the second missile clearly has a larger figure of merit than the first missile, suppose the derivation of

these figures were based on estimated enemy capabilities 3 years from now. If the second missile takes 5 or 6 years to develop, the validity of its figure of merit is doubtful. In this case, it would be necessary to re-estimate the enemy's capabilities at the time when it seems technically feasible to complete development of the second missile. From these new estimates, a new survival probability may be computed and a valid comparison made with the first missile.

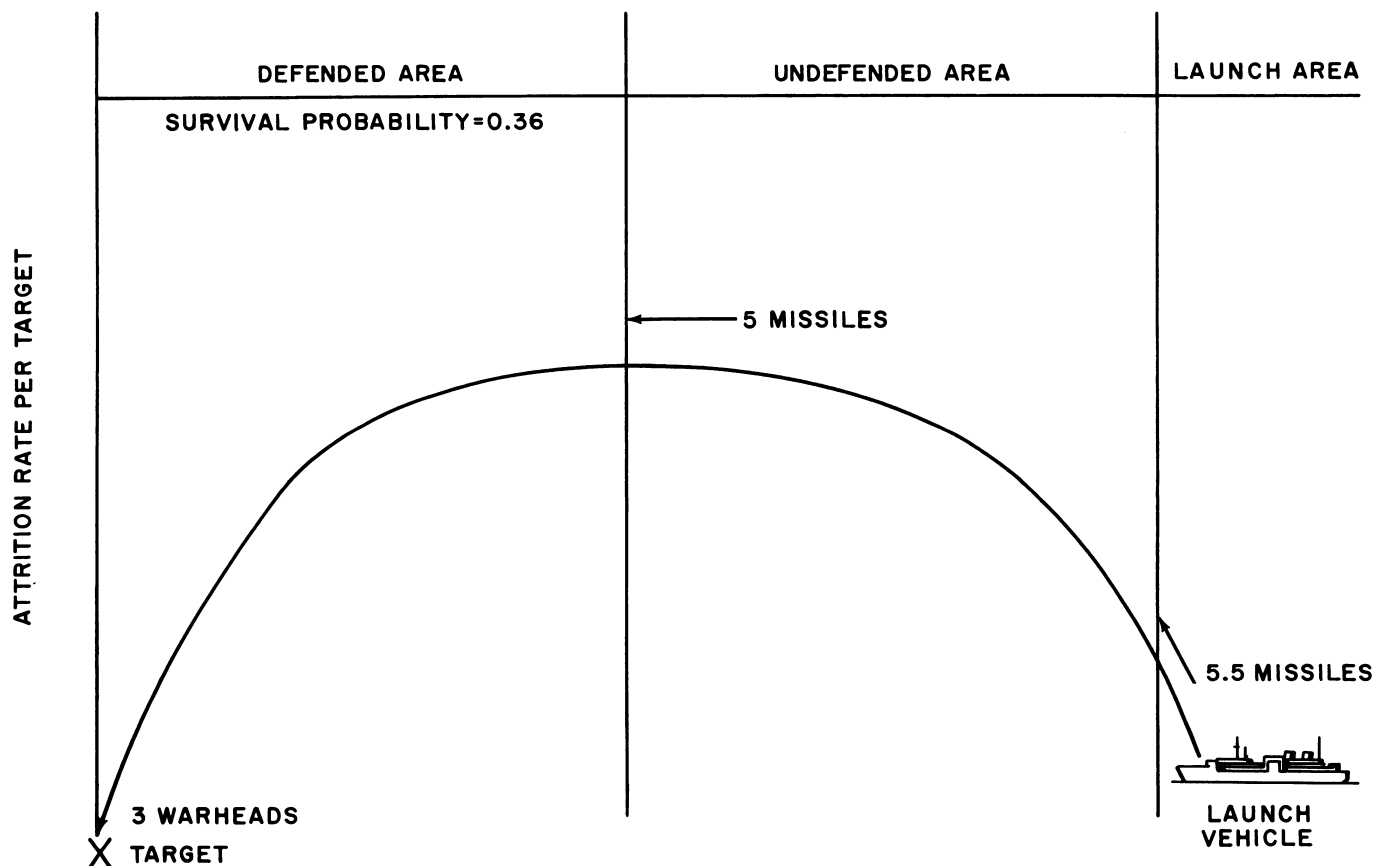
WEAPON SYSTEM EFFECTIVENESS

To determine whether there is an operational requirement for delivering nuclear weapons by guided missile when carrier aircraft are already available for this purpose requires a comparison of the effectiveness, that is ratio of dollar damage to the enemy to dollar cost for us, of both systems. In the case of missile costs, the survival probability is taken into account to determine how many missiles must be successfully fired to allow the required amount to reach the targets. Also, the number of warheads required to destroy each target must be accounted for. For example, if 9 targets are to be attacked and 3 warheads are required to destroy each target, a total of 27 missiles must survive the enemy defense. If the first missile is used (survival probability approximately 0.36) approximately 75 missiles must be

successfully launched. If we assume that 5 missiles will be prelaunch duds, a total of 80 missiles will be required to perform this attack.

Using the unit cost per missile and multiplying by 80 the total cost of the missiles can be arrived at. To this, various other costs such as a prorated share of the development and logistics costs for the missile, personnel losses, etc., must be added to determine the total cost of destroying the target(s).

The total dollar damage to the enemy may then be calculated and the effectiveness of the missile delivery method can be computed. In the same manner, the total cost of the carrier aircraft attack may be determined and the effectiveness of the system calculated. In this manner, the effectiveness of both systems may be compared and a decision as to whether or not to issue a new operational requirement may be logically determined.



FUNDAMENTAL CONSIDERATIONS IN WEAPON SYSTEM DESIGN

standardization of inputs

The point at which standardization of input takes place, and the degree of standardization of inputs, are two of the more important decisions which must be made at an early stage in the development of a system. In weapon systems, standardization of inputs such as various types of target information (i.e., type, quantity, position, velocity, etc.) may be necessary. It may be desirable to have

a highly sophisticated code at the logical control in the center of the system with a lesser amount of standardization in the manner in which the forward observer reports. Also, to account for unforeseen or improbable events, it is usually desirable to allow for some messages in unstandardized form.

group versus local optimum

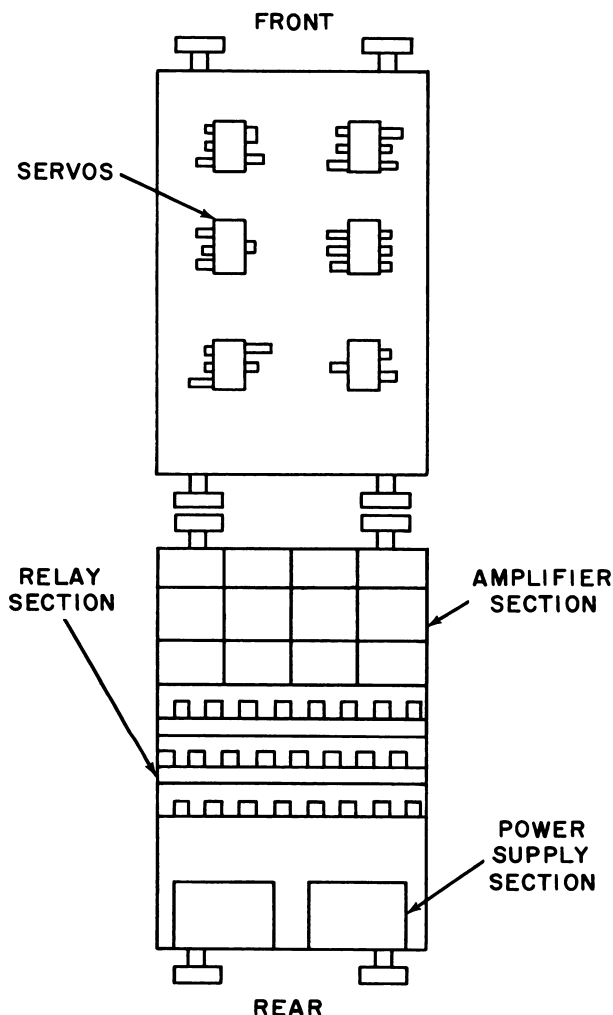
In order to optimize system changes, the entire system must be considered rather than certain elements. It is fallacious to assume that optimizing each element of the system individually will result in an optimized sys-

tem. Suboptimization also does not necessarily improve system performance. System design is the process of obtaining a group optimum.

sectionalization

The methods by which a system is organized into interacting subsystems depends primarily on the purpose of the subdivision. The breakdown of a system into subsystems which are functionally distinct is useful in analyzing system performance. However, it may be desirable to package together in groups those components requiring a similar technology or to arrange the groups in such a manner as to minimize interaction between groups. For instance in an analog computer, the servo packages might be packaged on the front part of the computing cabinet or servo section, and the amplifiers, relays, etc., might be packaged on the rear of the cabinet or electronic section. Although the functions with which the various servo packages are associated may be completely different and unique, all the servo packages require similar assembly techniques, handling equipment, test equipment, troubleshooting techniques, repair and replacement procedures, alignment procedures, spare parts, and personnel training. Design of a complex system in a manner such that separate sections containing components associated with a similar or identical technology are used is sometimes called the echelon design or sectionalized design approach.

When these sections are properly designed and built, each section may be replaced by a spare without requiring extensive alignment procedures to assure proper operation of the overall device. Thus, repair of the device consists of locating the malfunctioning section and replacing it with an identical spare section. In this manner the system can be made operational again with a minimum loss of time for repair. Once the overall system is again operating properly, the faulty section itself may be repaired independently of the system.



interaction between system elements

The strong interaction between elements of a system is one of the prime considerations in designing complex systems. For instance, the guidance and the control system of a missile must both act together in bringing the missile to target intercept before the target reaches its bomb release point.

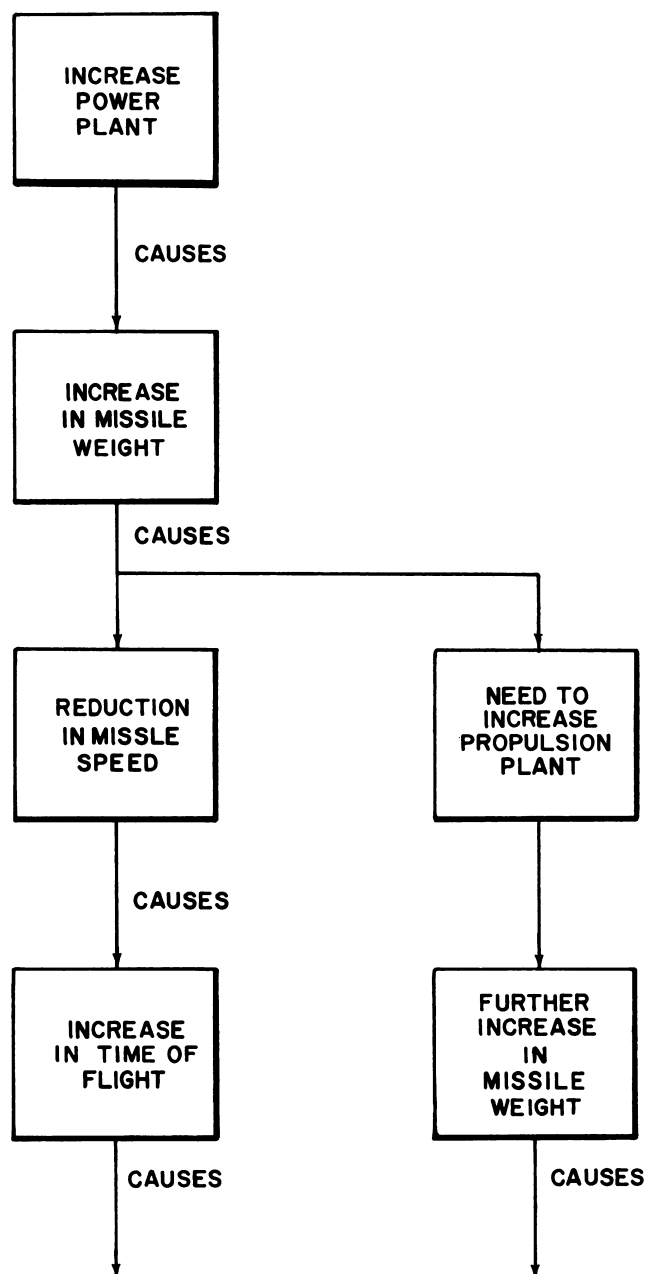
The accuracy of the guidance system is dependent to some extent on the performance and characteristics of the target. The accuracy of the guidance system is also dependent on the search and tracking radars which provide the positional information on which guidance is based. The characteristics of the airborne guidance equipment also affect the guidance system.

The accuracy of the guidance system in an engagement is therefore dependent on the combination of target characteristics and tactics, target location intelligence, and airborne guidance equipment characteristics. The closeness of miss is determined by the accuracy of the guidance intelligence, missile maneuverability, and target evasive tactics. The required closeness of approach of the missile to which the guidance system must operate is determined by the lethality of the warhead and the vulnerability of the target. The size and the power of the propulsion system is dictated by the weight carried by the missile as well as the required missile speed and the required maximum missile range. The weight of the missile is dependent on the equipment size, power plant size, maneuverability requirements, etc. From the above discussion, the interaction of various subsystems and system elements in an overall weapon system can be readily seen. Also, the importance of these interactions and their effects on the design of the overall system is apparent. For instance, the accuracy of the guidance system may be relaxed if a larger and more lethal warhead is used. Of course, a larger propulsion plant would then be required to maintain the same speed because of the increase in weight. Perhaps the speed requirement may be relaxed so that a new propulsion plant is not needed. Many more examples of the effects of interactions on system design may be derived to show the kinds of compromises which the system designer must make.

To provide further illustration of the problems involved because of the interaction between system elements, a specific system design problem for a guided missile is discussed in the following paragraphs.

If in a given design the electrical power supply is inadequate, the logical solution would be to increase the capability of the power plant. However, this would increase the weight of the missile, reducing the speed and range. A full study of all actions and interactions is obviously warranted. Possible solutions include:

- 1) Redesign of the electrical power supply to provide the necessary power with no increase in weight.
- 2) Redesign of the electrical equipment to use less power and thus make the power supply adequate.
- 3) Redesign of the propulsion system to compensate for the additional weight.



It is interesting to note that a problem involving a power supply may be solved by various approaches which may involve any major subsystem of the missile, including such subsystems as missile frame and structure, missile propulsion systems, and even missile warhead. In a major system such as a guided missile system, problems must be approached from the systems viewpoint. It is no longer valid to limit a power supply problem to the power supply specialist. The systems engineer has to create the evaluation technique for comparing various possible solutions on a common basis. Of course, the systems engineer must be assisted by the various subsystems specialists, but the approach to the solution must be from the overall viewpoint and not the specialized.

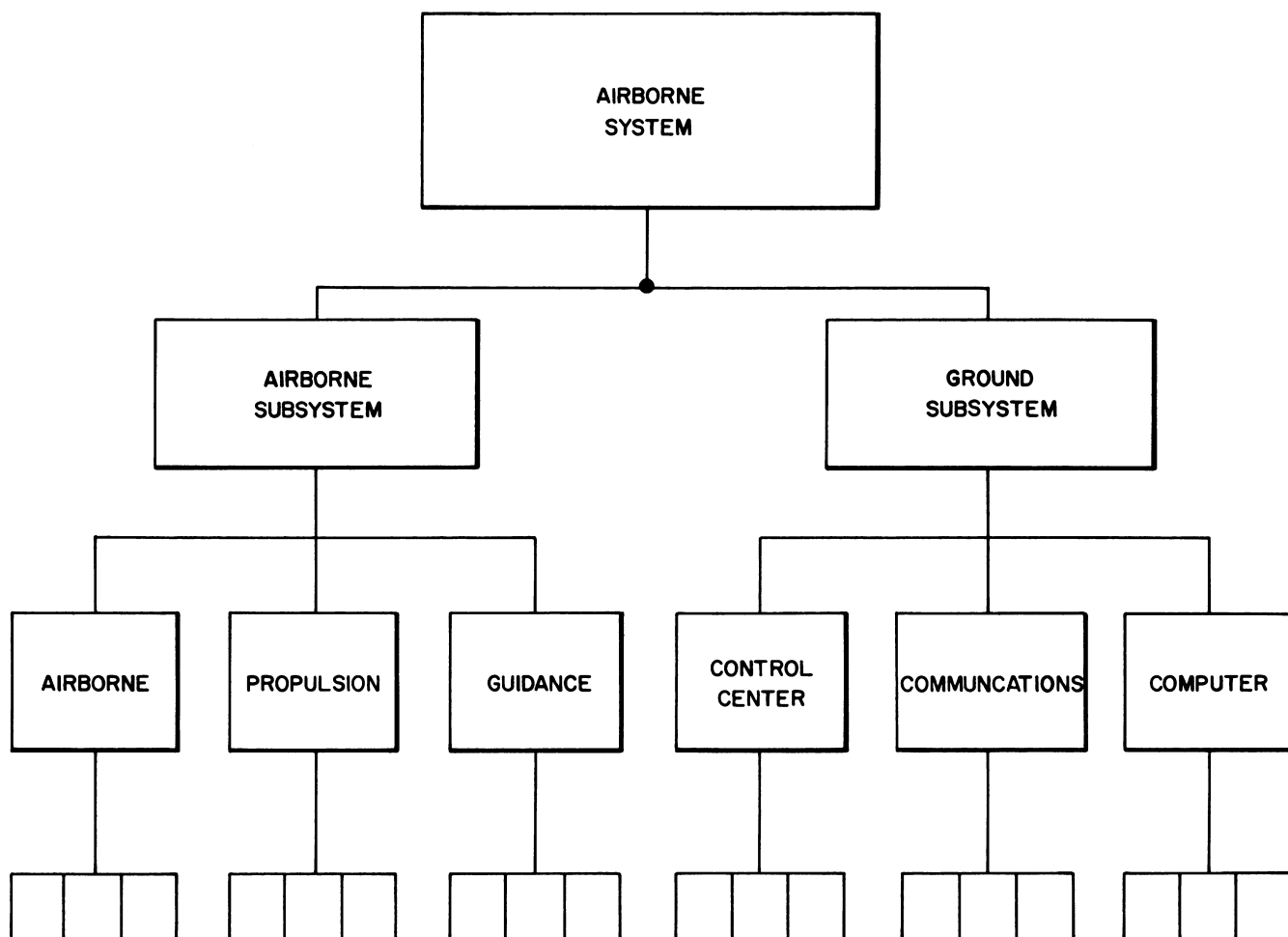
subsystem criteria

The choice of subsystems depends on a multitude of factors. One such factor is the desirability of having a subsystem in a single location. Another is that the subsystem should have as few inputs and outputs as possible. In the interest of reduced design and development time, subsystems should be defined so that an operating group can handle the design task without excessive discussion and conference time being expended on the relationship with other subsystems.

An example of the choice of subsystems is given by the subsystem breakdown of a digital computer. The com-

puter has control, arithmetic, memory, and input-output subsystems. The input and output functions are combined because their problems are similar and frequently they use the same equipment.

In most airborne systems, it is desirable to separate the airborne portions from the ground portions of the system. Each of these subsystems may be further broken down into lower order subsystems. For instance, the airborne subsystem may be broken down into air-frame, propulsion, and guidance subsystems. These subsystems are by no means independent of each other, but, with periodic coordination, their developments may proceed in parallel, more or less independently.



network systems

Almost every system involves a flow network of electronic signals or material. Communication, transportation, and power distribution systems are examples of network systems. Much work has been done on the mathematical analysis of networks to provide tools for design and analysis of network systems. Basically there are two methods with which to analyze a network. The first method consists of analyzing the path of each

and every signal or quantity of material passing through the network. The second method consists of analyzing the flow through intersections.

Since the number of intersections is less than the number of individual paths, the second method will be easier for complex networks. However, since it does not analyze the paths, it may be inefficient to transfer material on the paths. For some cases, it is then necessary to analyze the paths involved.

centralization versus decentralization

All decisions are made at a central command station and are based on information received from various elements of a centralized system. This contrasts with a decentralized system in which each element has the authority to make its own decisions and then report back to the central command station. In the decentralized system, it is understood that routine decisions will be made at the elements or lower echelons. Decisions of a more important nature should be made at a higher echelon; the more important the decision, the higher the echelon. In the case of an urgent and nonroutine situation requiring a decision, the lower echelon has the requirement to weigh the cost of making the wrong decision against the delay involved in referring the decision to a higher echelon. Since the lower echelon may not

always weigh the costs correctly and some wrong decisions may result, the centralized system has obvious advantages over the decentralized one.

The real distinction between centralized and decentralized systems is that communication between lower and higher echelons must be faster in the centralized system. In the decentralized system, the central command station maintains some authority over local operations based on the reports received from these operations over a period of time. The ability to exercise control of the situation throughout the entire system is therefore relative to the speed of communications. The more rapid the communications, the more centralized the system. There are four major disadvantages to the centralized system:

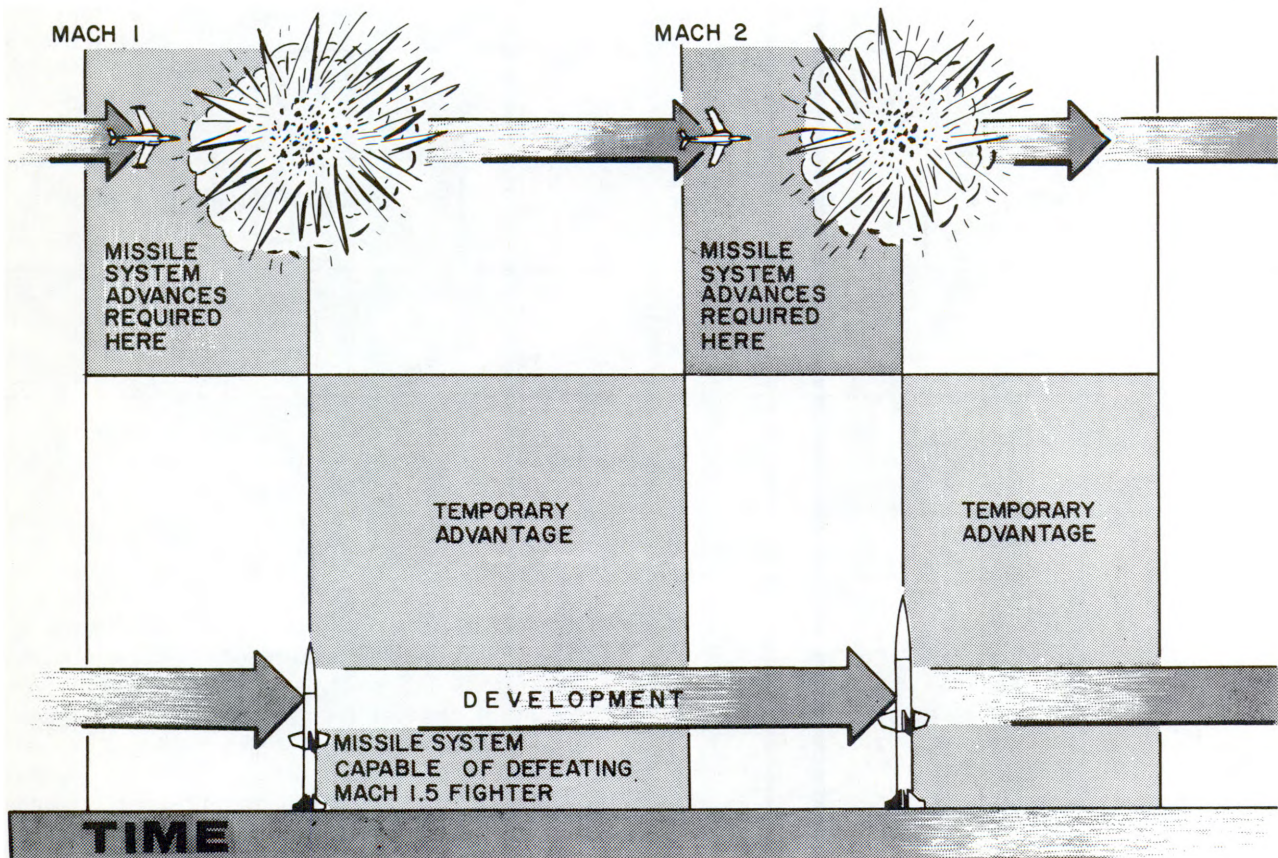
- 1) Inputs may be handled unsympathetically or slowly.
- 2) Information may be slightly colored or changed in its passage through the intermediate echelons.

state of the art considerations

One of the fundamental considerations in designing a new system is the time factor. To realize an advantage from a new, superior system, it must be produced before a potential enemy produces an equally good or better system. This consideration establishes a primary objective of systems engineering, that is, the production

of a superior system in minimum time.

The production of a superior system before a potential enemy or competitor can match it affords only a temporary advantage, since the competitor's technology will soon allow him to produce an equal or better system. Therefore, in order to maintain the initial superiority afforded by the first system, advanced versions of the system must be produced as early as possible.



3) Central command may be oversaturated with inputs and consequently may breakdown as a result of overload.
4) The system is inflexible; lower echelons become inoperative if their communications link with the central command station is broken.

However, the use of modern techniques such as high-speed, large capacity data processing units, large capacity data storage units (memories), high-speed computers, and modern communications techniques help to overcome these disadvantages.

Coloring of information may be avoided by employing reliable communications using codes such as pulse modulation and by employing multiple addresses whereby information proceeding up through the echelons goes to the immediate supervisor and simultaneously up to central headquarters. In this manner, the higher echelons get complete, rapid, and accurate reports on the operation of the entire organization while the supervisor of

each lower echelon organization retains his authority over the operation of his organization.

The use of multiple addressing increases the amount of information input to higher headquarters, possibly oversaturating its capability for handling information. However, with modern technology, if the system logic was well planned, mere multiplicity of inputs should cause no insoluble problems in data processing.

When a failure in communications between echelons occurs, each echelon must be able to operate without instructions from a higher echelon. If communications are interrupted or time does not permit communications with higher echelons, the lower echelons can take command and make decisions. Providing each echelon with the capability of command naturally causes this system to operate less effectively than the centralized system, but, at worst, it operates at least as efficiently as the decentralized system. Essentially, this plan insures that the system suffers no loss by being designed as a centralized system.

The above discussion serves to bring up a most important and critical problem in systems engineering: the conflict between performance objectives and time scales. Some of the factors which come into play in this problem are listed below:

- 1) A system is of no value until it is put into use.
- 2) Development of a new system usually requires a long period of time between its conception and use in the field.
- 3) The length of time required for this development can not be predicted accurately because it depends on the successful solution of many problems, some of which require varying degrees of invention.
- 4) The rapid changes in the state of the art tend to make systems obsolete during long development periods.
- 5) To survive competition, the system must simultaneously exceed in capability and precede in time any probable competitive system.

There are two choices available for producing a superior system. Each of these choices involves risk:

- 1) Develop a new system in a short period of time to avoid obsolescence. This can be done by using proven techniques and off-the-shelf equipment. The margin of superiority of this system over existing systems will probably be small, since no unique innovations or technical advances are utilized. Because of this, the margin of superiority may not be large enough to justify the costs involved in producing a new system. Also, the margin of superiority will disappear quickly due to the advent of new advances in the state of the art.
- 2) Attempt to develop a completely new system utilizing promising but unproven new techniques. The time and cost required for development and production of a system of this type will naturally be much greater than for the first type.

To summarize, it is obvious that the time factor is of great significance in selecting the method of approach to the problem. The performance sought and the time allowed must be optimized with regard to the overall superiority and advantage expected. The answer must come from a thorough analysis of the technological prospects of all major components of the system, the best obtainable intelligence of the possible action of the enemy, and the current state of affairs. A decision based on this analysis still involves a calculated risk but one in which the odds are heavily weighted on the side of success.

key system parameters

The three key system parameters that most often must be considered in system design are the system gain, the system time constant, and the system weight.

The system gain is a number which describes, for example, the ratio of the error in position of a guidance radar to the observed trajectory error.

The system time constant is a number which indicates, for example, the amount of time it takes the system to react to and correct a guidance radar position error. A time delay is frequently inserted deliberately to allow time to average out noise.

The system weight is classified as a system parameter since:

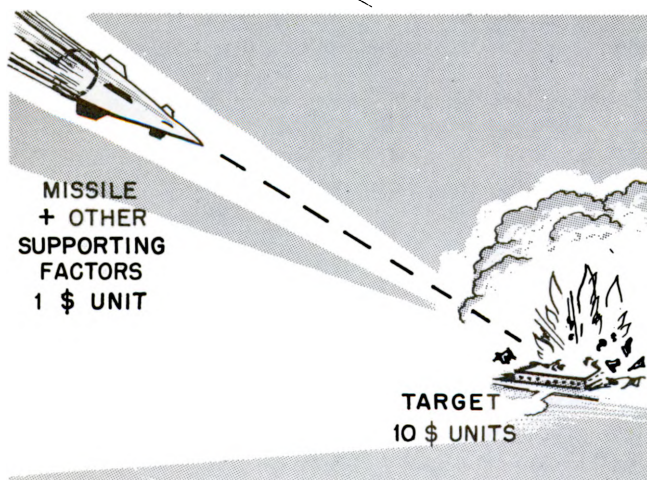
- 1) It is affected by all components, and
 - 2) It is basic to the applicability of a system to a particular location or environment, or, in the case of a missile, has much to do with the tactical performance of the vehicle in terms of maximum speed and altitude.
- Other key performance parameters such as cost, kill probability, accuracy, and development time also play an important part in system design.

measure of effectiveness

A measure of effectiveness must have many vital characteristics in order to be meaningful. It must be capable of being expressed as a number with unique meaning. It must be statistically efficient. It must have comparatively small variance and consequently be capable of being determined with reasonable accuracy and without excessive cost or delay. Other desirable characteristics are completeness and simplicity.

Certain measures of effectiveness in military systems come up again and again. Cost is almost invariably involved in these measures. For instance the effectiveness of a weapon is a figure of merit derived by dividing the total damage inflicted on the enemy in terms of dollars by the total cost of the weapon. In other words, a measure of effectiveness of 10 indicates that for every dollar expended, 10 dollars worth of damage is inflicted on the enemy.

Kill probability is also an important factor in measuring the effectiveness of a weapon system. Not only are we interested in the damage that the weapon may inflict but in the probability that in a given engagement, it will inflict that damage. A low kill probability requires a large number of missiles. This increases the cost per engagement, increases the length of time required to destroy the target, provides the enemy with more time to destroy us, and decreases the number of targets that can be handled by a given weapon.



alternative methods

The number of alternative methods available for system design is extremely large. In approaching the design of a new system it is necessary to analyze all of the methods and alternatives available and to make a choice based on the results of this analysis. The cost of developing and producing modern weapons systems obviate the luxury of making mistakes of omission, i.e., of not analyzing all the available possibilities.

morphological boxes

The use of morphological boxes is a system of thinking which helps the designer to visualize and analyze all possible solutions without regard or reference to standards of value. By viewing many aspects of a complex problem as a whole instead of following one aspect at a time, he can uncover new possibilities for solutions which might remain hidden to unorganized thought.

To set up a system in morphological boxes:

- 1) Write down all possible solutions to a problem with no initial prejudice about the problem outcome, but with a firm conviction at the outset that all solutions can be realized.
- 2) In a simplified way, make an estimate of the performance of each solution as follows:
 - a) State the problem to be solved exactly.
 - b) Determine the characteristic parameters, upon which the solution depends.
 - c) For each parameter there are a number of independent, irreducible values. List these in the form of a matrix.

P_1^1	P_1^2	P_1^3	$P_1^{k_1}$
P_2^1	P_2^2	P_2^3	$P_2^{k_2}$
P_3^1	P_3^2	P_3^3	$P_3^{k_3}$
P_n^1	P_n^2	P_n^3	$P_n^{k_n}$

NOTE: This construction is called a morphological box. Each of the $P_n^{k_n}$ values is a solution to the P_n^{th} parameter. Selecting one element from each row and joining these elements together as a system represents a possible solution.

- d) Determine by simplified analysis the performance values of all the derived solutions.
- e) Choose particularly desirable special solutions, and construct models to test them.

Although simple, the method appears long and tedious when one considers all the different analyses that must be made. The important idea, however, is to write down all the solutions so that they may be visualized together and so that none will be overlooked. Even though only a small portion of all the possible combinations is analyzed, much has been accomplished merely by writing them all down. It is possible in many cases to eliminate very quickly some values of the parameters in order to reduce the possibilities. In other cases only a few of the possibilities are applicable to the system task formulated. It is imperative at this stage not to discard unlikely possibilities too hastily just because they are radically different from the present trend.

To illustrate the use of the morphological method, consider some of the possibilities for use in a propulsion system. A morphological box for jet engines is:

$P_1^1 \quad P_1^2$	= intrinsic or extrinsic chemically active mass
---------------------	---

$P_2^1 \quad P_2^2$	= internal or external thrust generation
---------------------	--

$P_3^1 \quad P_3^2 \quad P_3^3$	= intrinsic, extrinsic, or zero thrust augmentation
---------------------------------	---

$P_4^1 \quad P_4^2$	= internal or external thrust augmentation
---------------------	--

$P_5^1 \quad P_5^2$	= positive or negative (suction) jets
---------------------	---------------------------------------

$P_6^1 \quad P_6^2 \quad P_6^3 \quad P_6^4$	= nature of conversion of chemical energy (thermochemical, electrochemical, radiation, or direct mechanical conversion)
---	---

$P_7^1 \quad P_7^2 \quad P_7^3 \quad P_7^4$	= vacuum, air, water, earth
---	-----------------------------

$P_8^1 \quad P_8^2 \quad P_8^3 \quad P_8$	= translatory, rotary, oscillatory, or no motion
---	--

$P_9^1 \quad P_9^2 \quad P_9^3$	= gaseous, liquid, or solid propellant
---------------------------------	--

$P_{10}^1 \quad P_{10}^2$	= continuous or intermittent
---------------------------	------------------------------

$P_{11}^1 \quad P_{11}^2$	= self igniting (hypergolic) or non-self-igniting propellants.
---------------------------	--

Detailed analysis of many of these solutions have been made. Examples of existing types of engines are described in the following paragraphs based on the morphological method.

RAMJET

$$P_1^1 \quad P_2^1 \quad P_3^1 \quad P_4^1 \quad P_5^1 \quad P_6^1 \quad P_7^2 \quad P_8^1 \quad P_9^2 \quad P_{10}^1 \quad P_{11}^2$$

or

$$P_9^3$$

TURBOJET

$$P_1^1 \quad P_2^1 \quad P_3^1 \quad P_4^1 \quad P_5^1 \quad P_6^1 \quad P_7^2 \quad P_8^2 \quad P_9^2 \quad P_{10}^1 \quad P_{11}^2$$

PULSEJET

$$P_9^1$$

$$P_1^1 \quad P_2^2 \quad P_3^3 \quad P_5^1 \quad P_6^1 \quad P_7^2 \quad P_8^1 \quad \text{or} \quad P_{10}^2 \quad P_{11}^2$$

$$P_9^2$$

LIQUID FUEL ROCKET

$$P_9^1$$

$$P_1^1 \quad P_2^1 \quad P_3^3 \quad P_5^1 \quad P_6^1 \quad \text{or} \quad P_8^1 \quad P_9^2 \quad P_{10}^1 \quad P_{11}^1$$

$$P_7^2$$

SOLID FUEL ROCKET

$$P_1^1 \quad P_2^1 \quad P_3^3 \quad P_5^1 \quad P_6^1 \quad P_7^1 \quad P_8^1 \quad P_9 \quad P_{10}^1 \quad P_{11}^1$$

While the use of morphological boxes does not directly help in arriving at the specific solution, it does ensure that the best solution is not overlooked. By writing all possible solutions for the propulsion system, we are not aided in the specific design of an engine. However, once the specific system is chosen, the principle of morphological thinking can again be applied to the subsystems.

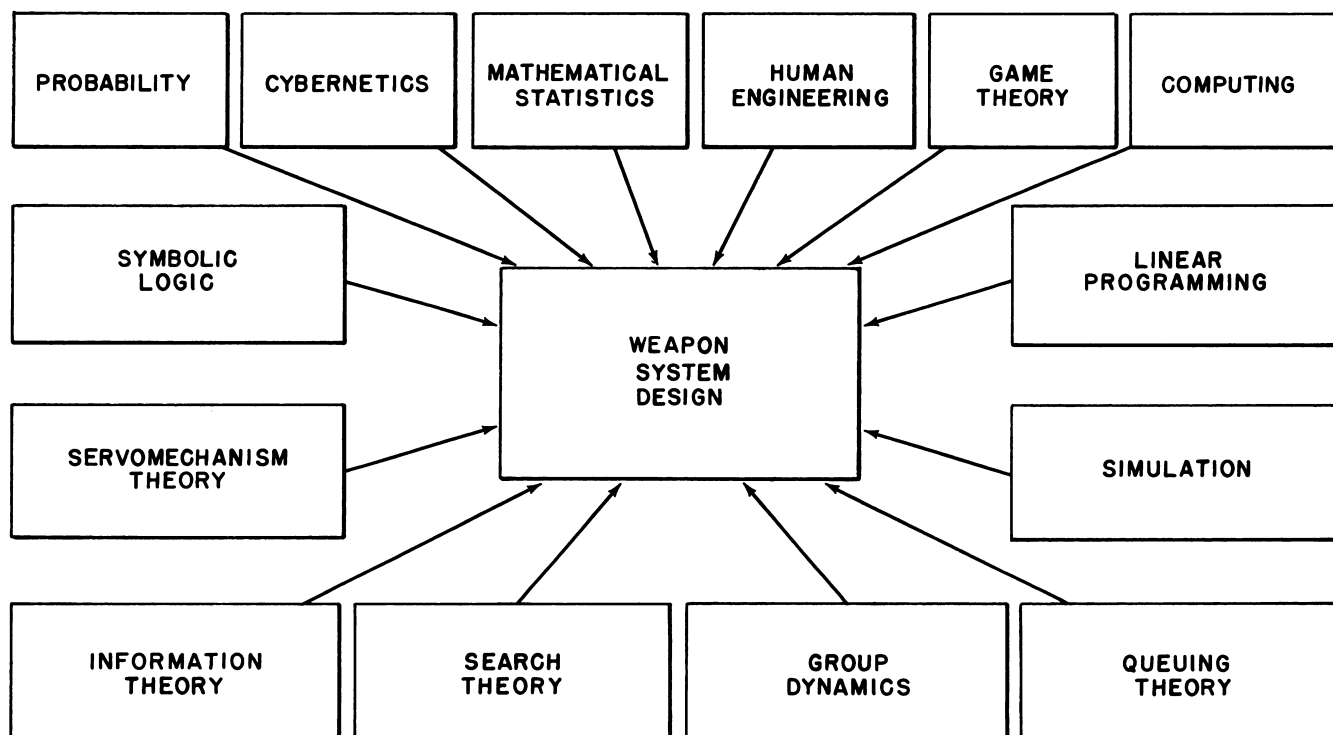
Economic feasibility

For a system to be economically feasible the materials required to produce it must be available in sufficient quantity and the cost of these materials must be compatible with the effectiveness of the system and with the nation's defense budget. Similarly, the system must be easy to manufacture, since the costs of manufacture affect the system in the same manner as the cost of materials. If the system costs a great deal of money to produce, it may not be economically feasible. Even if it is economically feasible, the prohibitive cost of manufacture may prevent large numbers of systems from being produced. The application of the system is thus limited. Because of its very nature, the rifle must be economically feasible in large quantities to have any value. However, an extremely large thermonuclear bomb may be economically feasible despite high costs per unit because it is intended for use against limited targets only. In this case, of course, the measure of effectiveness (damage to cost ratio) plays a great part in determining the desirability of the weapon.

METHODOLOGIES

-tools of system design

The design of a modern weapons system makes use of information ranging over wide areas of the pure and applied sciences, as shown in the illustration. Some of the more pertinent areas are discussed in the following paragraphs.



mathematical statistics

The field of statistics is concerned with the study of sets of observations and the formation of conclusions based on this study. Examples of the use of statistics in the design of weapon systems are as follows:

- 1) Quality control analysis embracing a realistic study of the variations encountered and the tolerances required throughout the entire weapons system
- 2) Analysis of reliability of systems, components, and ancillary equipment under the environmental conditions which may be encountered
- 3) Assessment of weapon system performance with respect to items such as single-shot kill probabilities, radar detection characteristics, etc.
- 4) Weapon system logistics
- 5) Study of development objectives.

Statistical analysis involves the systematic and scientific study of experimental data taken from a limited number of trials, with a view toward predicting future behavior, thereby verifying or extending previous conclusions reached on the basis of theoretical analysis only.

game theory

Game theory deals with strategies used in repeated competition such as small unit battles in warfare. The practical problems to which game theory is applicable are those in which there are conflicts of interests and the participants have some control over the outcome. In military systems, there are many situations which can be attacked by using a game-theoretical approach. Some of these are:

- 1) A submarine's attempting to remain undetected by the patrol plane
- 2) Bomber versus interceptor
- 3) The choice of a simple cryptographic code versus a complex one
- 4) The whole question of measure, countermeasure, and counter-countermeasure.

The objective of game theory is to find an optimum strategy, i.e., a strategy which gives the player the greatest expected value of payoff, maximum gains or minimum losses, and to determine the value of the game.

linear programing

Linear programing, a recently developed mathematical technique, is similar to game theory, in that it must frequently be brought down to an abstract form and variables must be assigned constant values. As of now, no concepts in systems theory are directly applicable to linear programing. Its infrequent use can be attributed to the ease with which problems involving linear programing may be transformed into game theory problems.

queuing theory

Queuing theory is the statistical method of estimating the delays and the waiting lines that occur whenever service has to be provided in sequence for inputs arriving at a random rate. Queuing theory can be applied to the aircraft that are stacked up over an airfield, or to messages awaiting transmission in a communications system. It has also been extended to include priorities so that a priority input goes to the head of the line or a priority message is handled first in a communications center. Problems arising from the effect of random distribution of targets and multiple targets on a weapon system may be solved using queuing theory.

search theory

Search theory, which is another highly developed technique, was first referred to the allocation of effort in conducting a search for a submarine known to be in a general area. The theory is used in military systems for antisubmarine warfare and antiaircraft warfare phases of defense.

information theory

There are two major aspects of the subject called information theory. The first concerns the quantitative definitions of the amount of information conveyed in a message and of the capacity of the communications

channel to transmit information. The theory may be used to optimize the design of a communications system. In general, implementation of information theory involves translation of messages to be transmitted into a form which is better matched to the communication channel. This aspect is usually referred to as the coding problem.

The second major aspect of information theory is the concept that communication problems are related to games involving the laws of chance or probability theory and therefore that a study of random processes can serve as a background to the study of communications systems.

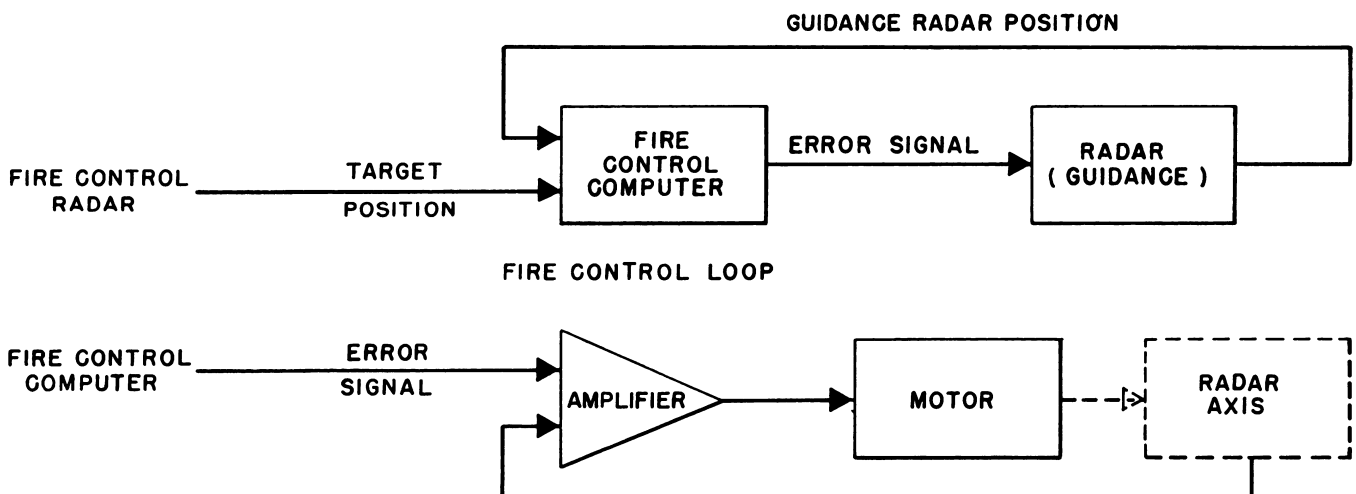
servomechanism theory

The heart of every automatic system is a control device of some sort. In large scale systems, this control can be in the form of logical control or reflexive control. Logical control is usually exercised by an automatic computer. Reflexive control is found in the form of servomechanisms. A typical basic servomechanism is illustrated.

A command or signal is fed into an error measuring device which also receives a measurement from the servo output. The resulting error between the input command and output measurement drives the servo to achieve a null condition. This null condition exists when the error signal has been reduced to the minimum, but not necessarily to a zero value.

A servomechanism may be used to control a radar position by slewing the radar to the position called for by the fire control computer. In this case the servomechanism is used in two places: first it is used in control internal to the radar; second, it is used in the entire fire control loop, where the fire control computer acts as the logical error-measuring device.

Servomechanisms are the basic elements in electro-mechanical analog computers. They are used as input-output conversion devices, computing devices, integrating devices, etc., within computers.



symbolic logic

The system of symbolic logic has not been applied universally to the field of operational research. However, its use in the design and evaluation of computers is invaluable. One of the basic rules in logic is that the logic is two-functioned, i.e., a statement may be either true or false. This logic is easily applied to digital computers since electronic devices are basically two-functioned. For instance, a switch is on or off, a transistor is cut off or is saturated, a signal is present or not present.

computers

The use of large scale, high-speed computers in the solution of problems arising in both the design and development of weapon systems and in weapons systems themselves has facilitated the rapid increase in weapon system capability.

One use of computers is to simulate the response of the weapon system. The weapon system provides outputs which can be mathematically defined in response to given input data. The system can then be set up on a computer as a series of equations and be tested to see if the response of the system is always what is desired. Once the design requirements are specified, the weapon system and its components are then designed to give the response which was assumed in simulation. In the areas of design and development, computers are used to optimize design factors, control schedules and continuously evaluate progress on development and production work, and, of course, to solve problems which allow new innovations to be applied to weapon system design. In the actual weapon systems themselves, computers provide the means to solve high-speed, multiple-target fire control problems automatically and accurately. They make possible the solution of problems in fractions of a second where normally many hours or days would be required. The advancing technology of computers continuously increases their capacity and speed while reducing their size.

cybernetics

The science of cybernetics is based on two fundamental concepts. The first is that all systems, living and mechanical, are information systems; the second is that all systems, living and mechanical, are feedback (i.e., servo) systems. In its broader terms, cybernetics includes all of the design of systems plus the understanding of systems, both mechanical and living. In the following paragraphs, samples of parallelisms of man and machine used in the field of cybernetics are provided.

NEURONS AND GATE

The most important means of communication within the organism is by means of the nerve cell or neuron. When the neuron carries a message, it is said to fire, i.e., some sort of electromechanical reaction starting at one end of the neuron is propagated down the neuron to its other end. Near the end of the neuron are a number of junctions called synapses at which the stimuli which fire the neuron are given. Thus, the combination of neuron and synapses may be considered analogous to a gate with a complex function of logical gate properties such as AND, OR, and NOT.

MEMORY AND LEARNING

While the nature of the human memory function is not known, several methods may be conceived which conform to a pattern compatible with neuronal mechanism. Two of these are circulating storage and threshold modification.

Circulating storage would be analogous to the acoustic delay loop line in a digital computer, where data is inserted in the loop, compressed, and recirculated indefinitely. Threshold modification would mean either that the stimulus required to fire a neuron was altered or that the synapses would be given different weights firing it. The second method would help to explain the conditioned reflex.

A reflex mechanism is one in which action is taken in response to an external stimulus without reference to the higher control centers. In the mechanical system, this means response by servomechanisms without communication with logical control, and usually it means action by the lower echelon without waiting for a decision from a higher echelon. In the organism, a reflex action is one controlled by a lower nerve center, the spine or the brainstem, without reference to the cerebral (conscious) portion of the brain.

GROUP DYNAMICS

Group dynamics had its origin in man's seeking to devise mathematical representations of the relations of groups of individuals to each other and their effects on each other. The basic application of the study of group dynamics at present is to communications systems. Group dynamics concerns itself with the efficiency of communications but, unlike information theory, it is concerned with the nature of the communication channels.

HUMAN ENGINEERING

Human engineering is generally concerned with ways of designing machines, operations, and work environment in a manner which capitalizes on the capabilities of human behavior and minimizes the effects of human limitations. The business of the human engineer is the engineering of machinery for human use and the engineering of human tasks for operating machines. In evaluating human mistakes the human engineer raises the questions:

- 1) Is the blame to be found in the design of the equipment which people use?
- 2) Do people make more mistakes on some kinds of equipment than on others?
- 3) Is it possible to redesign equipment so that human errors are reduced?

Research of the past 20 years has indicated that the answer to all these questions is yes. Based on this, the human engineer concentrates on making the equipment as mistake-free as possible by creating more easily readable and understandable control panels and displays, simpler and more flawless operating routines, and environmental conditions for the human operator that are conducive to more attentiveness and effectiveness on the operator's part. This involves the design of man-machine systems so that the tasks assigned to the man fall within his effective bandwidth.

SIMULATION

During the early stages of the development of a complex weapon system, the performance of the design being considered is normally computed through the use of general-purpose computers. As the development of the system progresses, however, certain portions or subsystems become available in the actual hardware form. By

using these portions of hardware for the analogous circuits or equations of the computation, a more accurate response of the system can be computed. If any undesirable differences between the desired output and new response are found because of variations between the response of the hardware and its assumed response, they can be remedied without waiting until the entire system is built.

SYSTEMS ENGINEERING

The systems engineer is responsible for:

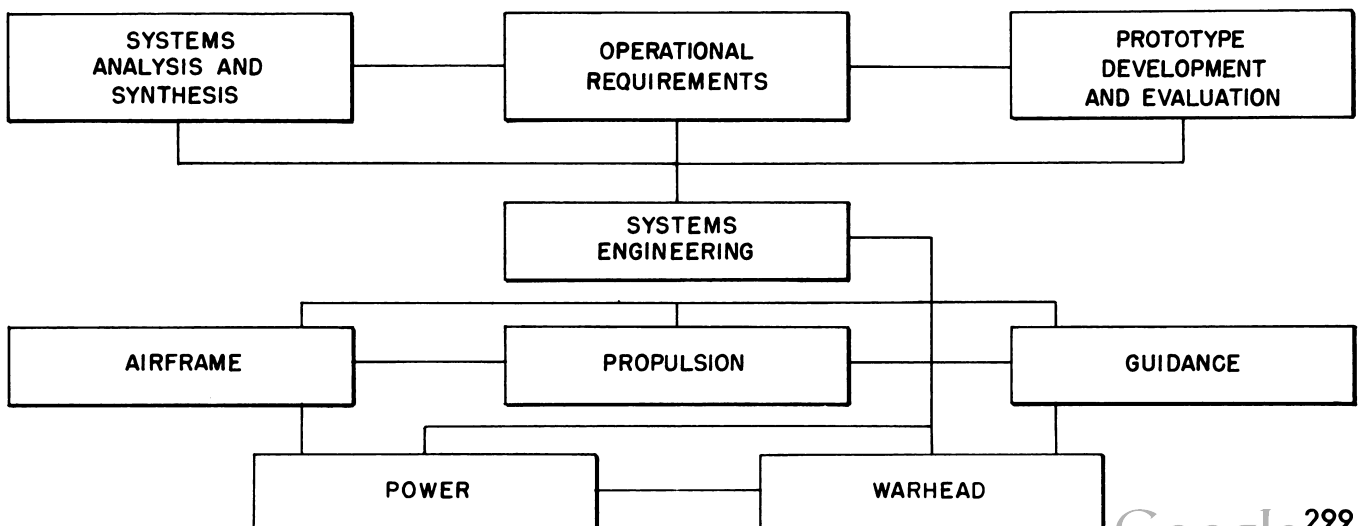
- 1) The formulation of an optimized concept of a system that will perform its function automatically
- 2) The translation of the concept into reliable practical hardware giving the desired performance.

Modern systems engineering implies the direct application of wide areas of scientific and engineering knowledge. It also implies that the function to be performed is highly complex, requiring selection, discrimination, and a certain degree of adaptability on the part of the system.

The problem of systems engineering can be broken down basically into two major areas. Both of these areas may be further broken down into subareas. The first area is concerned with studies of systems analysis and synthesis, operational requirements, and prototype development and evaluation. This area deals with the interaction of components and equipments in the system to be built. The selection of the equipment to be built is directly concerned with component interaction and is one of the systems engineer's most important functions. The second area is concerned with the technical activity leading to the solution of problems concerned with the design of components and equipment. For a particular missile system these problem areas may be the airframe, propulsion system, guidance and control systems, power system, and warhead.

Perhaps the responsibilities of the systems engineer may be summed up in five words: concept, selection, analysis, synthesis and evaluation. Once the function required is determined through operations research and eventual issuance of performance specifications, the initial function of the systems engineer is to formulate a

systems concept in which the structure of the whole system is outlined, its general methods of operation are determined, and the external functions of the various elements are defined. This calls for creativity, imagination, and broad background. During the evolution of a system, the process of selection is repeated by use. It must be exercised in the choice of one concept or another, in the choice of equipment, and in the choice of details of the equipment. Because of the interaction of subsystems, equipment, and components within a system this solution must be based on a thorough understanding of the overall system as well as a detailed knowledge of the subsystems from which the selection may be made. Critical thinking based on extensive knowledge is the basis of successful selection. The ultimate basis for selection is a thorough analysis which reduces opinions to quantitative terms and follows the scientific logic of cause and effect throughout the maze of the entire system. Analysis predicts ideal performance and also the departures from the ideal caused by the limitations of practically realizable equipment. A thorough and accurate analysis provides a scientific basis for design criteria, specifications, and tests. The integration of equipment into a theoretical system and the translation of these results into practical hardware that gives the required performance and can be produced, maintained, and operated by human agents, is defined as synthesis. Evaluation of a system tells how closely the practical hardware meets the requirement and provides the basis for the decisions on how the system can be used tactically. In addition, evaluation provides the data which will be used to design future systems.



organization and management

Once a detailed plan is established for attacking the various problems in the development of a system, organization must be developed to effect this plan through the prototype development. There are a number of different schools of thought on the method and manner of this organization, however, two of these are predominant in industrial engineering organizations: the departmental form and the task force form of organization.

In the departmental form of organization, the engineering department is composed of a number of semipermanent groups staffed by specialists in the various fields of engineering. Typical examples of such groups might be aerodynamics, structures, microwave, circuit design, electronic packaging, and so on, depending on the fields in which the department operates. In addition to the engineering departments, there are other departments or groups whose purpose is to provide services to the engineering departments. Some of these services may be drafting, environmental test, field test, experimental shop, etc. In this type of organization, projects are handled by assignment of a project engineer who reports either to a chief engineer or to a chief project engineer. The project engineer has the responsibility of coordinating the engineering activities carried out within the engineering departments. Assistance to the project engineer is provided by a small staff which is responsible for technical liaison, planning, scheduling, and monitoring of program costs. In the task force type of organization, the engineering department is organized entirely in terms of projects, except for service groups, as mentioned previously. Each project has a number of groups reporting to it, depending on the requirements and magnitude of the project.

In order to choose between the two types of organizations the following considerations should be taken into account.

MAGNITUDE OF FORCE

The number of people involved in a modern weapon system development program is exceedingly large because of the complexity and size of modern weapons systems.

NECESSITY FOR CONTROL

Because of the huge size of the work force and the many simultaneous subprograms in operation, extreme care must be exercised in maintaining adequate control over the project. Allocation of manpower, design of components which interact closely with one another, adherence of all portions of the effort to the program schedule, and cost of the program must be kept under tight control. Without such control, there is likely to be considerable uncoordinated activity which, if allowed to get beyond reasonable limits, can have detrimental effects on the overall program.

NECESSITY FOR DECISIONS

The necessity for decisions which may arise from unforeseen difficulties or breakthroughs again points out the necessity for effective control. In major weapons systems, many decisions made within one group which apparently affect one component may actually affect the entire system because of the extensive amount of interaction inherent in a major system.

analogy to a military campaign

The process of engineering a complex weapon system is very similar to that of waging a military campaign. Both operations require a large, highly trained, organized body of persons conducting a connected series of operations to bring about the desired result. Initially, the comparison of the systems engineering operation with that of a military campaign may meet with intuitive objections. Scientific progress has been mainly achieved by the creative work of a few individuals rather than a disciplined mass of technicians.

However, in the actual process of engineering a system, new principles and knowledge are not being sought; the fund of existing knowledge is being applied to solve the problem. The same is true about a military campaign. The campaign must be waged utilizing the forces and weapons at hand.

The system engineering process uses a large number of technical teams to attack the various problems at hand. In a military campaign, the individual small units also operate as teams. However, the team is inherently an organization of a small number of individuals of equal rank and cannot be extended to as vast and diverse a system as the engineering of a modern large weapon system.

The decision-making process in systems engineering is also very similar to that in military situations. Since all of the factors are rarely ever known with any great degree of reliability, decisions always involve the factor of risk. Although most situations can be attacked from several angles, the manpower situation usually has the effect of limiting the number of ways available.

Unforeseen circumstances may require immediate decisions so that no loss in momentum on the project occurs. The decision-making process in systems engineering may be described as technical generalship where the system engineer himself may be described as the technical general. The decisions made by the systems engineer are thus in scope and level analogous to those of a high echelon military commander.

The following general principles of organization used in military campaigns seem to apply effectively to the organization of systems development.

MOBILIZATION

Major systems development requires that both the manpower and material selected to do the job be mobilized in much the same manner as in the military situations. This type of effort cannot readily be absorbed into a general-purpose organization such as the departmental organization. The military places all the resources necessary to do the job under the cognizance of a single commander.

AUTHORITY

It is necessary that the systems development program be placed under the supervision of a single individual. This individual will still operate under the framework established by management for this program, but he will have the authority for rendering technical and program decisions within the framework. Because of the tremendous complexity and size of the program under his supervision, he must be placed in a high position in the organizational structure so that his decisions will be followed by the organization at large.

STAFF

Because of the magnitude of the job under the program supervisor, it will be impossible for him to make all decisions necessary to realize the program successfully. The program supervisor must be equipped with a competent staff of systems engineers with the necessary depth of systems and components knowledge in all the fields encompassed by the program. This staff must work as a mixed team to formulate the decisions required to keep the program moving along at its scheduled pace.

TASK FORCES

The various groups which must work closely together throughout the course of a system development must be under the program supervisor in order to insure a coordinated, efficient effort. The organization must also be flexible enough so that forces are mobilized and demobilized according to a definite program which serves to make maximum use of the funds available to perform the job as well as of the skills of personnel available for the program.

SUPPORT FORCES

Centralized supporting groups should be available to the organization but should not be under the control of the program supervisor. This method allows for optimum use of all the common support facilities and specialists who can best work away from the fighting front. The analogy described in the previous paragraphs seems to suggest an organization which is a compromise between the departmental and task force type of organization. In actuality, this compromise type of organization is the one most frequently used in present

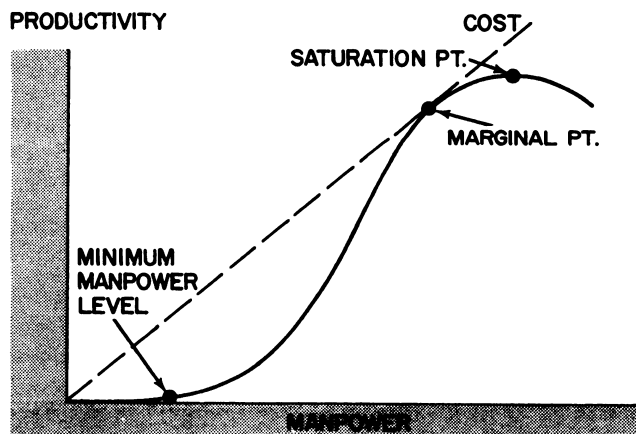
industrial engineering organizations. The resulting organization is closer to the task force type than to the departmental type in terms of the authority given to the man in charge of the program.

control of program effort

The necessity for exercising close control of the manpower engaged in a systems program cannot be over-emphasized. Because of the large number of persons involved, improper use of manpower will necessarily result in increased cost and time scales for the program. The major resource of the development organization is its pool of skilled technical manpower. The productivity and performance of the organization is directly proportional to the effectiveness with which this manpower pool is used.

A highly significant factor in the control program effort is the danger of overstaffing. The tendency to attempt to overcome lack of quality with sheer quantity often makes conditions considerably worse instead of better. It is fallacious to assume that the productivity of an organization always increases proportionally with the increase in manpower. In any organization for a particular program, there is a minimum amount of manpower required below which little or no productivity is achieved and above which productivity increases rapidly with an increase in manpower. This rapid rate of increase continues until the saturation point is reached. At this point only marginal increase in productivity results from increased manpower. Past the saturation point, increases in manpower either do not increase productivity at all or even decrease productivity. If the cost of the program were plotted on the same curve as the productivity, since both are a function of manpower, the curves would intersect before the productivity saturation point.

In other words, to design the program as economically as possible, the manpower assigned to the program should be less than the manpower required to produce peak productivity. If it is necessary to perform the job more quickly, manpower can be added to the program to bring in the schedule; however, the total cost of the program will necessarily increase.



control of component design

Because of the strong interaction between components of a subsystem, and the immense problem of assuring that all the components will eventually mate properly to produce the complete system as originally conceived and in accordance with all the performances specifications and program objectives, strong control must be exercised over the design, development, and production of the components of the system. Of necessity, the component designer must have a considerable amount of latitude in choosing his approach to the design problem at hand. However, there is a tendency to let this latitude assume proportions which may become dangerous to the realization of the program. There are three basic aspects to this problem which should be considered during the development program.

ALTERNATIVE DESIGN APPROACHES

There are always many alternative design approaches in the development of a system. As the amount of physical knowledge increases, the number of alternatives increase. With the current state of the art, this number becomes very large for many different applications. Each of the alternatives usually has advantages and disadvantages associated with it. The tendency is to develop several of these alternatives in parallel to determine which is best before making a final selection. Parallel development is used when the system is in the exploratory state, where the development and evaluation is done on paper or in the laboratory on a small scale, and when there is a major conflict involving achievable performance and potential risk between two alternative approaches. In most instances, if a single approach exists which will provide an adequate answer, additional development of other approaches usually turns up only marginal increases in performance and results in considerable unnecessary expenditure.

OVERDESIGN

The natural tendency of engineers to attempt to obtain performance from a component which is above and beyond what is needed to perform its function adequately and reliably is a problem very similar to that described for alternative design approaches. In a full-scale system development program, this type of effort results in additional expenditure and possibly in unnecessary risk if new and untried methods are employed.

DESIGN JEALOUSY

Since design tends to be an art rather than a science, evaluation of a new design by others tends to become subjective. In spite of this, it is necessary to provide well defined procedures for design review in order to provide collective judgment before the design is accepted for conversion into hardware. The systems engineering staff as well as others specially competent in various parts of the system should provide the required design review. The program supervisor should take part in this review so that he may maintain an up-to-date knowledge of the status and characteristics of the system. This is most important if the program supervisor is to be able to foresee difficulties and make corrective decisions with any degree of confidence.

control of program schedules

For the many reasons mentioned previously in this chapter, it is absolutely necessary to maintain control of program schedules if a successful program is to be achieved. This control, however, presents a complex and difficult problem.

While there are many programs where the initial schedule was maintained throughout the program and prompt delivery was achieved, the majority of programs do not fall into this category. It is more usual that a program slips substantially past the initially scheduled completion date. Surprisingly enough, there is a noted uniformity in the degree of program slippage.

The obvious question now is: why not add a safety factor to the initial estimate to take into account these unforeseen contingencies? This method is not a good one because of the tendency for organizations to utilize this time in improving the system or in pursuing insurance type programs. Therefore, all the time allotted for unforeseen contingencies becomes used up for other purposes, and when the contingencies arise, they result in approximately the same amount of slippage as before. Since the cost of development is generally a direct function of the time schedule, the longer the time allowed, the higher the cost. Unless the extra cost is compensated for by a substantial increase in performance over competitive systems, the additional cost becomes a net loss.

From the above discussion, the necessity of maintaining tight schedules becomes apparent. Extending the delivery date usually results in an increased slippage and certainly increases cost. This seems to imply that management must resign itself to the fact that schedules will slip. However, this is not always true. The entire problem of schedule control still remains unsolved and it is obvious that it is worthy of serious study. Currently, there are several new methods being employed to maintain control over schedules. These methods make use of computer facilities and advanced techniques. Basically, three directions in which savings in time may usually be achieved are:

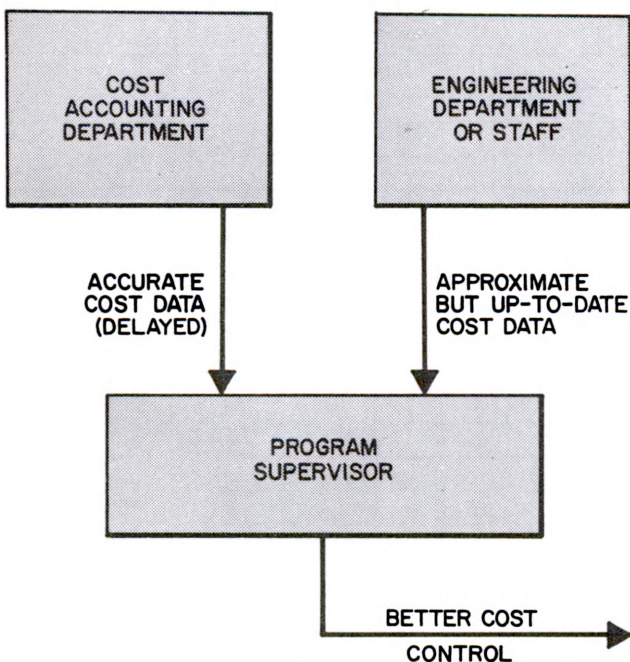
- 1) Simplify the objectives which the system was initially intended to meet. Very often, after analysis, some of these objectives are found to be extras which are not basic to the solution of the problem.
- 2) Simplify the method of attack by eliminating studies and applications of alternative design approaches for the purpose of optimizing a function once an adequate design approach is already available.
- 3) Apply the organization specifically to the job at hand, and alter the organization as required to improve on schedules.

control of program costs

The importance of control of program cost in a system development program has been repeatedly discussed throughout this chapter. There are three factors which are not always properly handled: maintenance of current budgets, authority for expenditure of funds, and cost consciousness.

MAINTENANCE OF CURRENT BUDGETS

While most organizations maintain a fairly elaborate cost accounting system which allows people to charge time to projects they are working on, the results of the tabulations of these charges often are delayed so that information on an exceptionally large cost does not become available to the program supervisor until it is too late to do anything about it. In maintaining control over cost, it is necessary that the program supervisor continuously have an approximate accounting of expenditures to date, so that he may evaluate the course of the remainder of the program against the remaining funds. This information is a valuable addition to the exact accounting afforded by the organization's overall cost accounting operation. In most cases, this additional up-to-date accounting must be done within the project staff or within the engineering department itself.



AUTHORITY FOR EXPENDITURE OF FUNDS

In many organizations, persons responsible for making high level program decisions within the organization have no control over funds spent outside of the organization. This can result in needless red tape which would be avoided if the program supervisor were given the authority to expend funds, provided they are within the established program's plan and budget. The organization management must still monitor these expenses, and exercise overriding control when deemed necessary.

COST CONSCIOUSNESS

In general, engineers do not have an inherent drive towards cost consciousness. Most engineers are more concerned with achieving high performance in their equipment than with striving for low cost. Very often, too little consideration is given to the use of design methods during the prototype development stage which will aid the eventual production problem. Since production engineers have little latitude to make changes once the basic design has been established at the prototype stage, this problem poses quite a bit of difficulty. One solution is to have production engineers working in every phase of system design and development. In addition, cost reduction programs must be initiated and pushed to develop cost consciousness among engineers and other technical people. New engineers should be oriented in the ways of production engineering so that prototype designs will take into account the eventual production problem.

prototype evaluation

The evaluation of a prototype is accomplished only after all the various subsystems and equipments have been mated together. At this time, it is possible to evaluate the performance of the system against the performance specifications and program objectives.

The evaluation of a prototype is a painstaking job, requiring a complete and exacting series of tests exercising all of the equipment and functions existing in the system. In most cases, these tests require a multitude of various test equipment, jigs, setup, tools, etc. For example, the evaluation of a prototype guided missile system consists of a long series of ground tests using signal generators to simulate guidance inputs, large-scale shake testing equipment, drop towers, and many other items of special test equipment to detect any system defects. After the laboratory checkout, it is necessary to carry out a series of prototype flight tests to demonstrate that the system operates in the same flight environment. The analysis and interpretation of ground and flight test data is in itself a major task and requires a highly trained analysis group.

The end result of the prototype evaluation is to determine if the original specifications are fully met so that any below-standard performance may be corrected prior to the initiation of the production run. In addition, the final decisions on the design of the system test equipment are made at this stage.

system production

The engineering of a final system begins when a commitment is made to produce the initial quantity of a system. This commitment may be made before or after prototype evaluation, depending on the degree of risk involved as compared with the pressure for an early completion of the task. In the case of a completely new and advanced system design, the first prototype may not meet the design objectives and a second and third prototype may be necessary.

The final engineering process represents detailed design of the individual components and sections, resulting in finished engineering drawings for production. Environmental factors, production factors, and reliability factors are all considered in detail in the final design. Final design of a system is the most unglamorous phase of systems engineering. However it is also frequently the most hazardous. Some of the more common pitfalls are weaknesses of detailed design, departure from prototype design, premature freeze, production factors, and complexity and operability.

CHANGES FROM PROTOTYPE DESIGN

During the time interval between the freeze on the prototype design and the start of the final engineering task, various advances in the field may indicate areas of improvement of the prototype design. However, since most new techniques have not been proved, it is usually impossible to predict whether they will work or not. Because of this, the temptation to change the prototype design should be resisted.

During the course of final design, there are usually many changes which must be made to the prototype design to meet design objectives. Any unnecessary changes only increase the risk. By making explicit provisions for the incorporation of such changes after the start of production, the equipment may be kept up to date in an orderly manner without undue risks.

THE PREMATURE FREEZE

A premature design freeze for the prototype may increase the desire to include new techniques in the final engineering phase. The process of final design is not well suited for making changes. To make a change during this phase requires changing many final engineering drawings as well as other types of documentation. Also, should a change affect other components as a result of the interaction within the system, a major group of changes may result. The design freeze date must be chosen using the highest level of technical insight and judgment so that premature freeze problems are avoided.

PRODUCTION RATES

Most modern complex weapon systems have specialized applications and therefore are produced in small quantities in the order of tens or possibly hundreds. Because the production run is small and mass production techniques may not be profitably used, the unit cost is usually high. To keep costs down in this type of operation, it is necessary to devise new production methods such as special multifunction tools, etc.

NUMBER OF PARTS

The number of parts used in a system is a direct function of the system design. In many cases, the system may be simplified by removing unnecessary features, thereby reducing the number of parts required. Of course, it is necessary to ascertain that the removal of parts does not place overloads on the remaining components, causing them to operate above their prescribed limits. The removal of parts without checking their interaction thoroughly can cause disastrous results. This is another reason why top-flight system engineers must be used during the final engineering phase of a system.

SPECIAL PRECISION COMPONENTS

Very often, the use of special precision components in a system ties up a considerable amount of the overall cost of the system. These items are usually custom built, very expensive, and are often delicate and particularly subject to environmental conditions. The tendency of the systems designer to impose special requirements for certain components causes these components to be custom built rather than standard off-the-shelf items. A method for standardizing such components so that they may be mass produced at lower cost is needed for the solution of the precision components problem.

COMPLEXITY AND OPERABILITY

Most systems tend to be complex because the functions they must perform are of such magnitude and complexity. As the complexity of systems increases, the number of components increases and hence, for components of fixed reliability, the reliability of the system decreases. Because of the larger and more complex systems presently in existence, extreme care must be taken with each component, no matter how small and trivial, in order to insure the required amount of system reliability.

Because of the complexity of modern systems, imaginative designs must be employed to allow for rapid troubleshooting and repair. Modular design, automatic built-in test equipment, strategically located test points, and easily used and understandable control and test panels are among the methods used by systems engineers to enhance maintenance of complicated systems.

SYSTEM OPTIMIZATION

ANALYSIS OF PERFORMANCE

Once the preliminary selection of components is made, an analysis must be made to determine the characteristics which will result in an optimum system. This problem amounts to the determination of a set of specifications on the inputs and outputs of the various pieces of hardware which make up a weapon system. The system block diagram is a useful aid to thinking at this stage of development. The block diagram is arranged with the subsystems represented by blocks with lines representing signal flow interconnecting these blocks. For instance, the tracking radar locks on the target and feeds information to the computers. The computer then determines the required launcher orientation and sends orders to the launcher, directing it to this position. If the missile is a beam rider, it must be fired into the guidance beam. Therefore, the computer also generates guidance radar positioning orders which it sends to the guidance radar. The weapon control system receives tracking information on all targets and the control officer selects the particular target to be engaged. After the launcher and guidance radar are positioned, the weapon control system transmits the fire order. On this order, the launching system fires the missile from the launcher so that it is captured in the beam of the guidance radar. Based on the continuous target position data received from the target tracking radar, and the continuous missile position data received from the guidance radar, the computer generates orders to the guidance radar which causes the missile to follow a prescribed trajectory to target intercept.

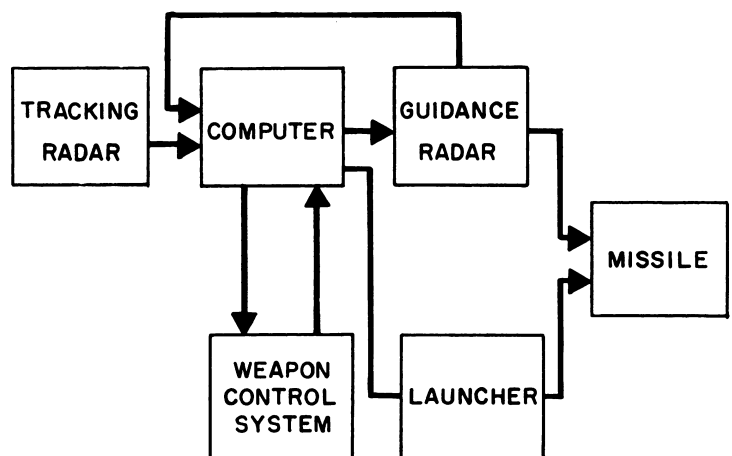
The system performance equations are determined by the system transfer functions, which relate the outputs of each of these blocks to the inputs. These outputs are not always related to the inputs alone but contain disturbances internally generated. Some of these sources of error are:

- Target scintillations
- Receiver noise
- Radar range limit
- Dead zones
- Gyro and amplifier drift
- Gyro crosstalk
- Receiver saturation
- Computer approximations
- Maximum computer accuracy

The problem then is to analyze system performance, in the presence of these disturbances, in terms of the possible transfer functions under our control. Since the components have not been constructed at this stage of development, there is a wide range of values of the disturbances which may exist and an equally wide variety of transfer functions which may be used. A very good way of proceeding with the solution of this problem is called the method of successive approximations. In this method, the systems analyst, after deliberation and consultation with hardware people, guesses at some probable values of the disturbances on each component. Then, he chooses a set of key system parameters which are assumed to describe the operation of the system. The key parameters usually chosen are system gain, system time constant and system weight, as was discussed previously in this chapter. After choosing the key system parameters, the analyst proceeds to determine the values of these parameters which produce an optimum system. In determining these values, the analyst uses any means or tools at his disposal, such as analysis, computers, etc. The optimum system is the system which gives best performance of the system task.

Performance must also be described in terms of performance parameters such as cost, accuracy, kill probability, and development time. Using the optimum values of the system parameters, tentative specifications, temporary designs, drawings, new estimates of the system are made. These new estimates plus some additional parameters are next introduced into the analysis. In the meantime, hardware construction is started. Tests on the development hardware are made and the results of these tests are fed back into the analysis.

It is often found in the first analysis that many of the system errors have relatively small effect on system performance, but that one or two are critical. In this case, it is desirable to reduce the magnitude of the critical errors before reoptimizing the second time around. When these errors are reduced as far as possible, only a few key system parameters and only a few critical errors are left.



TRANSLATION INTO FINAL DESIGN

The translation of the system design into hardware involves specifications and drawings. Specifications must be written in terms of the using organization rather than in terms of the functional units of the system. For instance, although for optimization purposes an aerodynamics system can be described as a transfer function for various speeds and altitudes, it must be translated into wing and tail dimensions, body diameter, structural loads, surface temperatures, alignments, etc. All electronics, whether tracking radar, guidance radar, computer, or missile electronics, may be developed by one group, while all electromechanical components are developed in another. Specifications are written in the languages of the respective groups involved. This involves compromises. For instance, how much of the gain or time constant of the system is to be

allowed to the electronics units and how much to the electromechanical units? All of these considerations must be studied and resolved, using all the tools available to the systems designer.

The optimized system is then translated into drawings and specifications and is given to the shop to manufacture. After the first hardware units are constructed, extensive tests are made on the individual components to determine accurately the performance of each. This data is then sent back to the systems designers, who proceed to correct their initial decisions as necessary and possibly even their initial assumptions. After the initial phases have been modified by the results of the hardware construction, new hardware may be built to the revised specification. In this manner, the system design converges to the optimum point aimed at by the initial design.

CHRONOLOGICAL PHASES

in naval weapon system development

operations research

The first phase in naval weapon systems development is operations research. Initially, operational requirements must be established. The capabilities and limitations of the weapons system are set forth. The development characteristics and development plan are also established. After these items are completed, the initial phase is terminated and the second phase, system design, is initiated.

system design

At the outset of the system design phase, performance specifications are established. For instance, the range at which the incoming target must be detected, the capability of the detector to distinguish targets from decoys, the accuracy requirements of the guidance system, the kill probability of the missile, etc., are all established. Also, compatibility requirements must be established to insure, for instance, that the missile system is capable of being installed aboard a particular class of vessel and that it is capable of operating properly in its intended environment.

Once these requirements have been set forth, preliminary system design is initiated. Models are built to mechanize the performance specifications. These models may be largely mathematical at first and may utilize computers, simulation techniques, and other modern technologies. As the design continues, components are developed and inserted into the model so that they may be analyzed and evaluated. This analysis is continuous and allows for checking system operation from the beginning to the end of the design phase.

As the design continues, experimental devices are fabricated and tested. Evaluation of these devices may indicate areas of redesign necessary to insure conformance with the specifications. The continuous testing and evaluation of new devices and of the overall system through use of the model allows for development of the overall system within the requirements of the performance specifications and compatibility requirements. The principle design of the system may then be accomplished with confidence and a fair amount of exactness so that the production phase of development may ensue.

production engineering

Prior to the start of a production program, production prototypes are built. These prototypes are produced in the same manner as the eventual production models. During the course of the prototype manufacture, deficiencies, or bugs may be detected. The detection of these deficiencies during the prototype phase allows for a smooth efficient production of the final equipment. Faults which show up during the production effort itself may prove to be expensive and probably will cause schedule slippage. Bugs which are detected in the prototype may require redesign to eliminate them. In many cases, the operation of the production line may be tested by pilot plant production to insure against faults in the production line itself. All of these operations and tests in the production engineering phase of system development are preparatory to the actual equipment production and allow for efficient and economical production as well as more reliable systems.

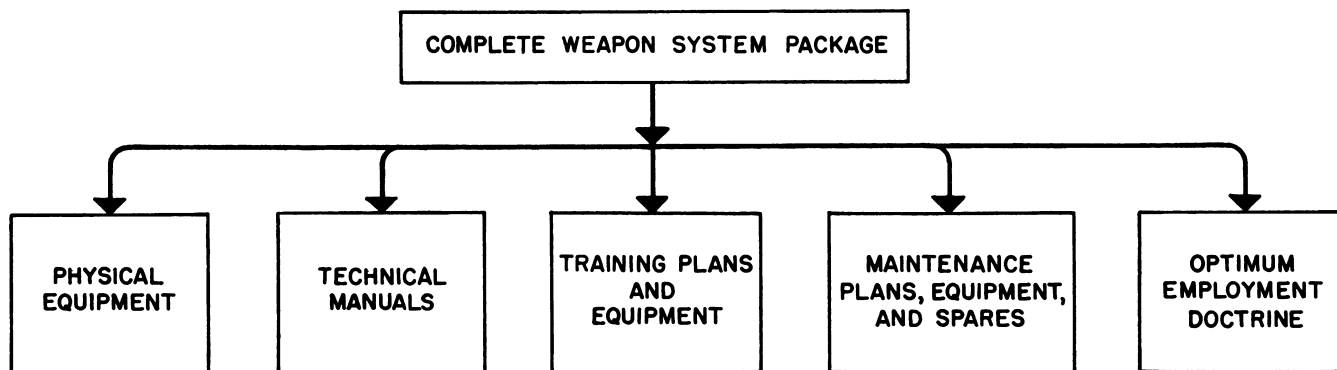
release to fleet

After the weapon system is produced, it is released to the fleet, where it undergoes further analysis and evaluation. This consists of BuWep's technical evaluation performed by engineers and technical per-

sonnel of the Bureau of Naval Weapons. Actual system operation in the fleet is accomplished by means of the OPTEVFOR evaluation. The system is then introduced to the fleet and proper training programs for naval personnel are carried out.

THE COMPLETE WEAPON SYSTEM

The complete weapon system does not consist of the physical equipment alone, but also of technical manuals, training plans and equipment, and spare parts, and of an optimum employment doctrine. Weapon system development thus includes concurrent development of all of these items. Planning, finding, and development must include these elements which must be completed by the time the weapon is ready to be introduced to the fleet.



Technical manuals are usually developed with the system prototype so that preliminary copies are available for review with the prototype. Upon completion of review and approval of the preliminary manuals, final manuals are prepared incorporating all of the changes made during the evaluation of the prototype. These manuals are then issued to the fleet along with the production equipments to serve as a guide for the proper operation and maintenance of the system. Manuals are kept up to date through the issuance of change sheets to cover changes to the system due to retrofits, ORDALTS, etc.

Training plans and equipment are also an essential part of the weapons system package. Adequate training is a must to ensure proper operation and maintenance of modern complex weapon systems. Training may consist of classroom lectures and studies as well as work with and on the actual systems. Training equipment may consist of models, analogies, charts, and the actual equipment itself. These equipments provide visual and physical demonstrations of the theory and operation of the actual system.

Maintenance plans provide procedures and intervals for

performing periodic, preventive, and corrective maintenance on the system. These plans are essential to minimize down time and increase the effectiveness and usability of the system. Test equipments are necessary to maintain modern complex systems. These equipments may range from voltmeters and oscilloscopes to specially built test sets which provide actual dynamic conditions and loads for system testing. Built-in equipment which automatically and continuously monitors system operation and pinpoints malfunctions is another feature of modern weapon systems.

Along with the equipment an adequate amount of spare parts must be furnished to allow for repairs of the system. The amount of spares is determined by the failure rate of the components, the amount of time for which the spares are required, and the repair philosophy established for the system.

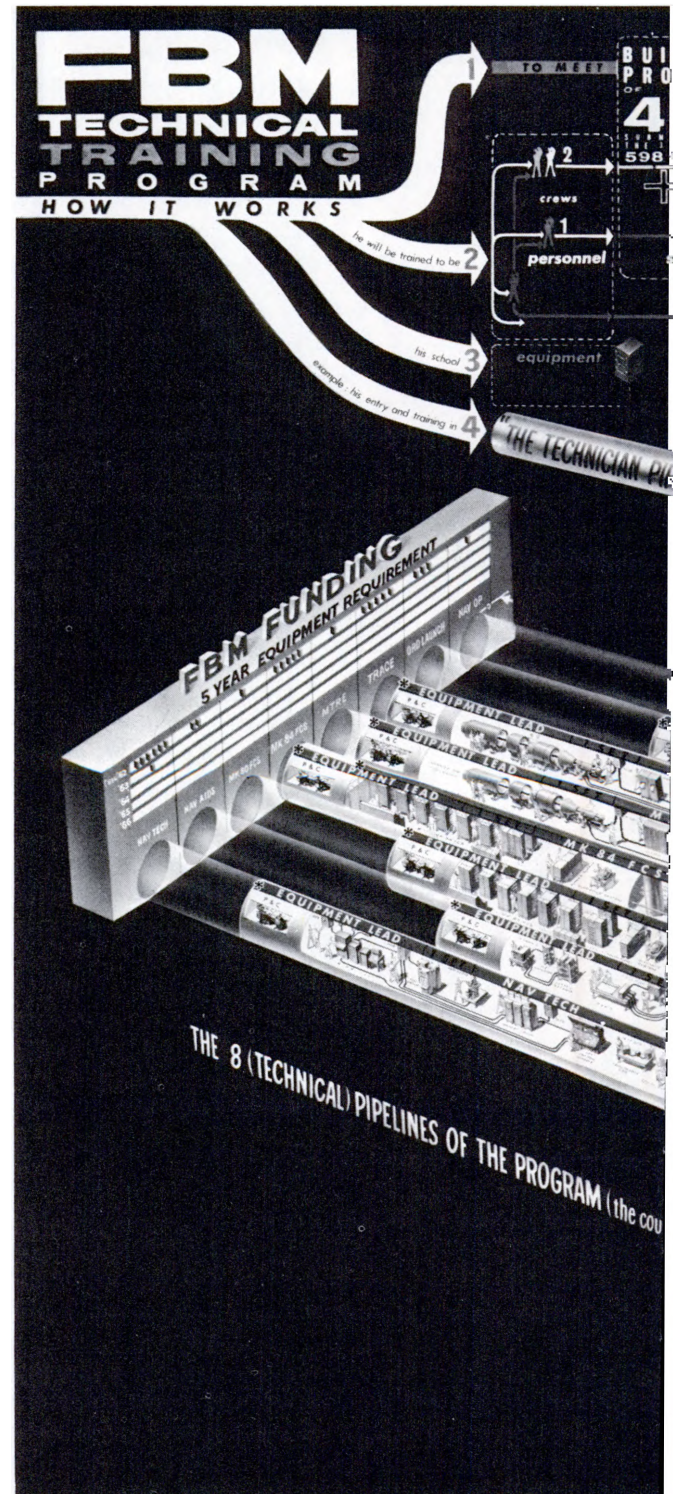
During the development stages of the system, studies are made to determine the optimum employment doctrine for the system. This doctrine must be completed with the issue of the system to the fleet. In this manner the equipment may be utilized in the fleet with maximum effectiveness.

training plans and equipment

In the planning stages of a well organized weapon system development program, there is need for an evaluation of the training requirements necessary to insure proper operation and maintenance of the physical equipment of the system. These requirements are based on the number of subsystems involved in the system, the eventual location of the system, and the personnel responsible for the operation and maintenance of the equipment. In the planning of the training program, consideration should be given to training necessary during development of the equipment, training during field evaluation, and training after delivery of the equipment. Training during development may be carried out by the training activity of the hardware contractor. This procedure varies with the individual contractors. In one case, the training activity assigns highly skilled engineers not directly associated with the design group to study the equipment and prepare a training format which will be utilized by the eventual users of the equipment in a formal instruction class. The engineers developing this format make use of the information generated by the publications activity of the contractor and in many cases bridge the gap between the information in an instruction book and the actual need of the personnel assigned as users of the equipment. This is possible because the training engineer acts as an instructor and is directly associated with the eventual crew.

In this training program, the crews are oriented in the basic functional theory of the equipment. They are trained in the actual operation of the equipment and, whenever possible, use actual prototype equipments for this study. The crew is introduced to and firmly grounded in the maintenance philosophy, both preventive and corrective, for the equipment. When the training is completed and the equipment is finally delivered, the contractor has no further obligations other than to furnish replacement parts for the equipment. Of course, this type of training is limited to small groups of men, usually that crew which is assigned to a unique system which will have only one or very few installations. When a large number of identical systems are built, the providing of schools and instructions becomes impractical for the manufacturer. In this case, the training is accomplished by assigning skilled personnel of the buyer the task of learning the equipment. These individuals then assume the responsibility of training lesser skilled personnel.

The training of entire crews may be facilitated by the use of equipments which simulate the operational equipment. Training devices are utilized to familiarize the crew members with the makeup and operation of their equipment and to teach them how to make more effective use of the equipment by providing them with an understanding of various conditions of operation. The time of the student is fully utilized because there is no waiting

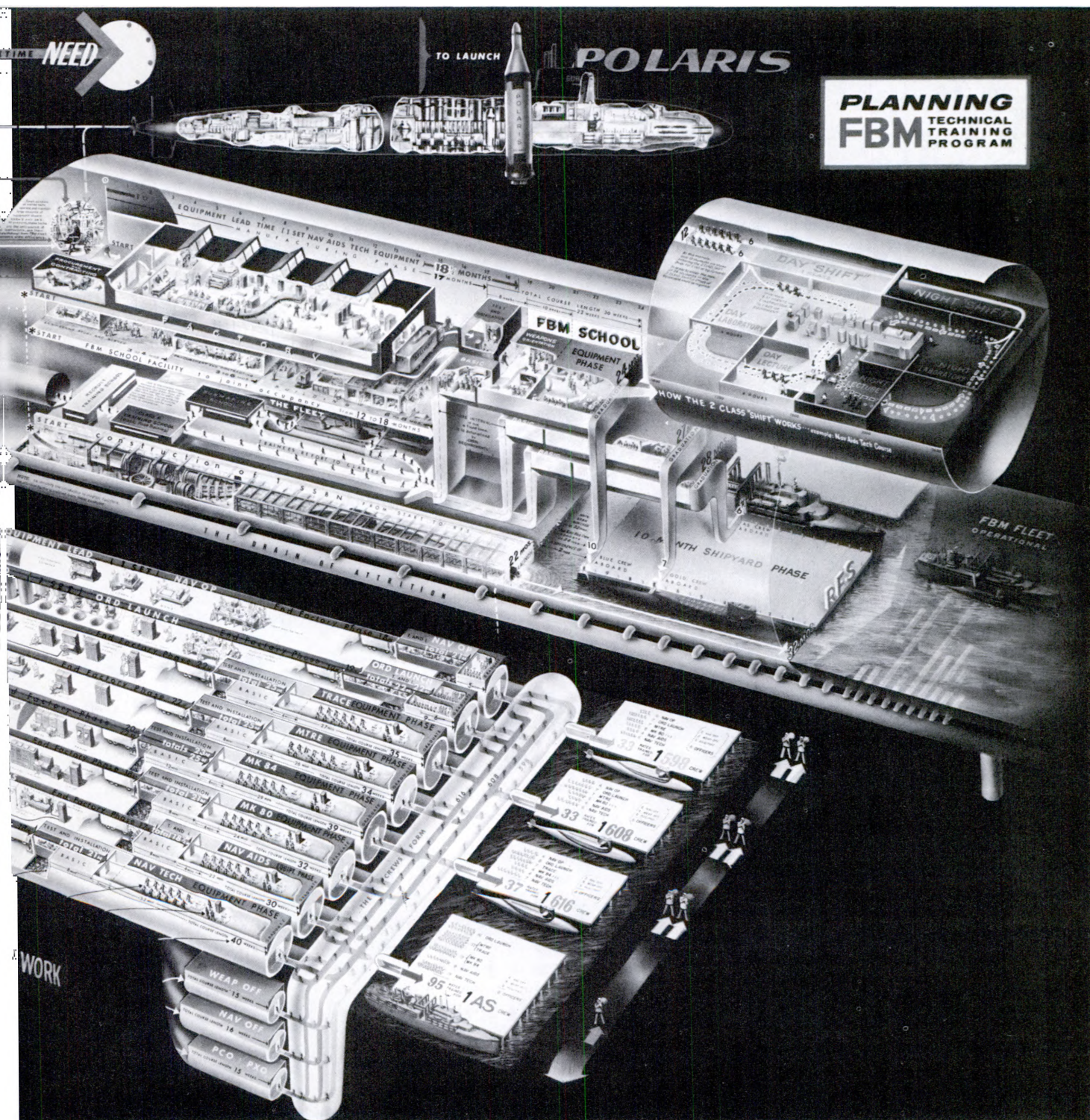


for availability of the equipment, supporting personnel and other necessities. Operating costs of these simulation equipments are a fraction of the cost incurred by the utilization of the actual equipment.

Equipment layout is such that the facilities for instruction are in general much better than in the operational equipment. Time spent in learning actual equipment is reduced. Hazardous emergency situations can

be simulated and recovery from them practiced, with no danger to either personnel or equipment.

In maintenance training, the conditions of work and practice provided by the training equipment are arranged for the greatest possible safety consistent with realistic operation. Safety habits acquired during such training are carried over to the operational equipment by the student.



The utilization of special training devices makes possible the effective development and coordination of the skills required of individuals and the integration of all crew members into a tactical team. Trainers bring about savings in checkout time, reduction in accidents, and increase in proficiency through increased technical knowledge of both normal and emergency procedures. In the training of crews a specific schedule is usually followed, consisting of:

- Problems based on the latest tactical doctrine
- Transition and refresher training
- Training in emergency procedures
- Training in normal procedures
- Training in instrument procedures
- Actual mission rehearsal
- Crew evaluation

In training based on the latest tactical doctrine, the students are brought along slowly. As student proficiency increases, problem complexity can be increased until all aspects of the tactical situation have been included.

Transition and refresher training is vital in the restoration of knowledge and skills that have deteriorated through lack of use.

Training in emergency procedures provides practice in recovery from simulated emergency situations. Practice immeasurably increases the probability that the operator will act correctly in an actual emergency. Training in normal procedures insures that all phases of normal operation are conducted realistically. All operations and procedures that are conducted with the operational equipment can be performed in the trainer while the instructor visually and aurally monitors them for compliance by all crew members, with their assigned checks and other duties.

When the entire crew is engaged in a tactical problem, most of the operation is conducted under normal conditions. Any failures or malfunctions inserted are limited to those which will not cause the practice mission to be aborted, since the object is to train the crewmen to act as a team in carrying out the mission. This continues until complete harmony is attained.

Some of the many devices utilized by the naval training activities are listed below:

- Aviation devices
- Basic science devices
- Communications devices
- Radar devices
- Navigation devices
- Armament devices
- Recognition training devices
- Surface (sea) operations devices
- Undersea (sub-surface) operations devices
- Tactics devices
- Teaching aids
- Films
- Automatic rater cards
- Charts and posters
- Instructional recordings
- Antisubmarine warfare devices
- General training devices
- Survival training devices

Many of these training devices are suitable for use directly in a classroom; others may be too large for this and are therefore more or less permanently installed at one or more of the many naval training stations located throughout the United States. These larger equipments are not readily suitable for shipboard use because of their size; however, devices of smaller size can be installed and used on shipboard.

maintenance plans, equipment, and spares

Effective maintenance is no accident but is a design factor in the development of weapons system equipment. From the initial conception of system designs to final production, development of maintenance equipment and maintenance philosophy should parallel the development of the equipment.

Every weapons system requires test and checkout equipment. This equipment can be designed as an integral portion of the system or can consist of external equipments, either of special design or standardized commercial items. The subsystem under design usually defines which type of test equipment (internal or external) the subsystem can utilize most effectively. For example, a computing element in the system would utilize an internal test and checkout equipment built directly into the subsystem. A test program for a computer can be established which, when run into the machine, can indicate the readiness for operation of the equipment. The receiver portion of a radar set would utilize a target video simulator, a specially built external equipment, for test and checkout. This equipment generates target video, the parameters of which simulate a target

which could be theoretically detected by the radar in normal operation.

Various types of meters may be used during the operation of a system to monitor self performance. For example, voltmeters are used to monitor the regulation of the many power supplies in the system. In the high-power transmitters found in radar sets there is need for close monitoring of the currents and voltages necessary to power the RF devices.

Some systems come equipped with built-in oscilloscopes which, in addition to displaying normal signals occurring within a system, can also be utilized during test to check out circuits suspected of malfunction.

Many analog computers containing servo loops and amplifiers use automatic amplifier failure detector circuits which are built in and continuously check the operation of each servo amplifier in the system. The circuits are such that when an amplifier malfunctions, it is immediately identified.

In digital computers, the addition of simple toggle switches and display lamps so that test data may be inserted and answers may be read out makes it possible to check the entire operation.

Many electronic equipments utilize so-called module packages of electronic gear to facilitate repair and replacement when malfunction occurs. These module packages are a direct result of miniturization of electronic devices. The transistor and printed circuitry techniques have enabled electronics manufacturers to produce equipments composed in large part of these modules. The manufacturer also supplies module spares. Modular units facilitate rapid replacement of a whole circuit very much like tube replacement in older electronic equipment. The difference is that the module usually contains an entire circuit. The advantage of these modules is that localization of a malfunction can be more rapidly performed, since it is no longer necessary to troubleshoot down to a vacuum tube or in the case of the module, a transistor. When the troubleshooting procedure localizes a malfunctioning module, it is immediately replaced, thereby reducing the time an equipment is inoperative.

Routine maintenance procedures are formulated to maintain the equipment in a state of optimum operation and also to prevent premature failure of the machine by monitoring normal operating level and by replacing system components suspected of pending failure. The routine procedures are preventive in nature and when properly carried out maintain the equipment in a continuous state of readiness. More money and time is saved by the application of effective routine maintenance procedures before the equipment breaks down than by the most detailed troubleshooting procedures applied after equipment failure.

Troubleshooting procedures are formulated to localize a malfunction in a system in the most efficient manner possible. These procedures are based on actual symptoms of failure, such as loss of sweep trace on a radar system PPI scope or an incorrect answer on a computer read-out display during checkout.

Although it is possible to include in a troubleshooting

chart all the symptoms of malfunctions which could occur in the complex systems of today, it is not recommended. A relatively small number of possible malfunctions have a high probability of occurrence. One disadvantage in the development of a troubleshooting philosophy is that it is necessary to assume only one malfunction at a time, although it is probable that at certain times more than one component may fail.

The final item in effective maintenance is the need for complete spare parts, since with them troubleshooting is brought to a successful end.

It is obvious that spare parts should be completed and ready when the weapon is introduced into the fleet and should be available with the weapon system and in the supply system when the weapon joins the fleet.

technical manuals

Technical manuals are an element of our defense structure that is frequently underemphasized. To produce effective military technical manuals, management understanding, capable personnel, careful planning, adequate funding, proper specifications, good plant facilities, engineering cooperation and coordination, and complete printing and distribution facilities are required. Technical manuals must be available on board a ship for each item that is to be supported. They must be made available from the moment the equipment is placed in operation to have full effectiveness.

Among the many problems to be solved in bringing about effective technical manual programs is the consolidation of the many specifications covering technical manuals for the services.

The exchanges of certain documents between contractors on a controlled basis is particularly helpful to avoid the pitfalls of incoordinated work in the development of manuals. An example of this is the problem of different symbol designations given to parts by different contractors. Early interchange of information among such producers of manuals is most essential. This eliminates confusion by standardizing symbols so that they may be easily understood by personnel making use of the manuals. The final test of any technical manual is that given it by the user in the field. The manual should be as nearly perfect as is feasible by the time it reaches its ultimate audience.

To assure this high degree of accuracy and effectiveness, the manual should undergo various checkouts and proofings along the way. The principles and certainly some of the practices which apply to quality control of hardware are also applicable to manual production.

Finally, there is need for a system to generate and promulgate quickly those changes in the equipment made necessary by discovery of errors, omissions, etc., and by changes to the hardware itself. The failure to provide such a system presents a very serious problem. Technical manuals are a vital part of a weapon system program, demanding a high degree of capability in their production. When they are produced in an effective manner, they permit the full, effective, safe and economical use of the related equipment.

optimum employment doctrine

With the advancing technical pace of today, many weapons systems are on their way to obsolescence before they even reach the hands of the users. Therefore, while a system is in development there is need for review of present tactical operations and formulation of new tactical doctrines with regard to new weapons. This need is especially pointed in a review of the history of wars and the weapons of war. Spanning history, from the time of the Roman Empire to our own 20th Century, the time elapsing between the development and eventual use of superior weapons was determined largely by custom and tradition rather than by an intentional and systematic study of the advantages of using the superior weapon. Superior arms, if employed immediately and effectively, can bring about a victorious decision even before enemy countermeasures can be formulated. It follows that the development of new weapons and the doctrines concerning their employment have an important bearing upon the success or failure of a nation's military force.

In the past we have often been guilty of three failures: Failure to adopt, actively and positively, the thesis that superiority of arms favors victory.

Failure to recognize the importance of timely establishment of weapon employment doctrine.

Failure to devise effective techniques for recognizing and evaluating potential weapons as science and technology advances.

A new weapon may have a strong influence on naval strategy and fleet tactics. Therefore doctrines must be reevaluated and revised accordingly. This reevaluation of tactics must be completed well in advance of the actual reception of the new weapon by the fleet. The fleet must be firmly instructed in the optimum employment doctrine of the weapon before it arrives so that valuable time is not wasted in acquiring this knowledge. This is necessary since the weapon has a comparatively short useful life and combat requirements are urgent. Employment doctrine must be formulated concurrently with the development of the weapon so that the full potential of the weapon may be realized as soon as it joins the fleet. Methods of determining a weapon's optimum employment doctrine have been formulated. These methods consist of tactical analysis techniques, simulation, and tests of first production models. These tests are classified as operational suitability tests and tactical effectiveness tests.

Initially, the operational performance of a weapon is predicted by an analysis of design data and experimental tests. Once the weapon is produced in prototype, the

predicted operational performance characteristics must be confirmed and if necessary modified. To do this it is necessary to plan an evaluation test program of the weapon under actual operating conditions. Thus the operational suitability of the weapon will be proved and data will be collected from which the best tactical doctrine can be derived.

In determining the operational suitability of a weapon system, it is necessary to prove the compatibility of the system with the expected physical environment, the ease of operation of the system by field personnel of average qualifications and the compatibility of the system with already existing and associated weapons and material. As an example, consider an air-to-air missile system. This system must be compatible with the weather, altitude, temperature, humidity, vibration, and landing shocks imposed upon it by the parent aircraft, as well as by the physical environment of its own flight. The crews which operate the system will not, of course, be specially selected, but will have received specialized training. The scope and direction of learning should be observed. Incompatibilities between the missile system and associated weapons and material should be noted.

The determination of tactical effectiveness is performed by a planned test which is broken down into naturally sequential parts. The individual probabilities of success for each part can then be determined as functions of the tactical variables which are significant. Finally, the individual probabilities corresponding to a particular tactical situation can be multiplied together to get the overall probability of success for that situation. Overall tactical effectiveness can be determined by examining the probabilities of success for the expected tactical situations.

The net result of a well planned operational evaluation test program should be a handbook of tactical performance for the system at hand. Such a handbook should be used as a basis for tactical training in peacetime and for modification of tactical doctrine, if required by the realities of war.

Peacetime tactical doctrine should provide maximum training against the type of opposition expected from potential enemies. Accordingly, foreign intelligence should be combed for useful information. The most likely enemy tactics should be formulated. Each of these should then be analyzed, using the tactical performance handbook, to determine what combination of variables under our control will give best weapon system effectiveness. The results should be established as doctrine for peacetime training and initial use in war.

D217.12:
3000/v.3

NAVWEPS OP 3000 (VOLUME 3)

NAVWEPS OP 3000 (VOLUME 3) WEAPONS FUNDAMENTALS: SYNTHESIS OF SYSTEMS